3 MATERIALS AND METHODS

3.1 Experimental Procedures

3.1.1 Cell Culture

C2C12 Cells

Skeletal muscle development is a process involving a series of steps. First multipotential, mesodermal precursors are committed to a muscle cell fate. This is followed by myoblast proliferation, activation of muscle specific genes and fusion to form multinucleated muscle fibers[231]. The C2C12 murine myoblast cell line was first established by Yaffe & Saxel in 1977[232]. It provides a well established model to investigate this process, as myoblasts proliferate in undifferentiated state when cultivated subconfluently in media supplemented with 10% FBS. When cells reach confluence C2C12 cells rapidly differentiate and fuse to give multinucleate myotubes accompanied by cell cycle withdrawal. This process is further assisted by maintaining cells in differentiation media containing 2% horse serum instead of 10% FBS. The molecular marks of muscle differentiation are also observed in C2C12 differentiation in a chronologically appropriate order[233].

C2C12 cells were obtained from Prof. Jakob Schmidt (Department of Biochemistry and Cell Biology, State University of New York, Stony Brook, New York) and cultivated at 5% CO2 and 37 °C in Dulbecco's modified Eagle's medium supplemented with 1% Penicillin/Streptomycin and 10% fetal calf serum. Mononucleate myocytes were harvested at a level of less than 70% confluence. To induce differentiation, cells were cultured with Dulbecco's modified Eagle's medium and 2% horse serum and maintained for 48 h, by when more than 90% had fused into myotubes[234].

HL-1 Cells

HL-1 cells are a cardiac muscle cell line, derived from the AT-1 mouse atrial cardiomyocyte tumor lineage by Claycomb *et al.*[235]. HL-1 cells can be serially passaged, yet maintain the ability to contract and retain differentiated cardiac morphological, biochemical and electrophysical properties, maintaining a cardiac-specific phenotype. As is typical of healthy mitotically active cardiac muscle cells, the cytoplasm of HL-1 cells is filled with nascent myofibrils and areas rich in glycogen. HL-1 cell generally contain one centrally located nucleus surrounded by contracting myofibrils. Cells posses perinuclear atrial natretic factor (ANF) containing specific granules, cardiac-specific myosin and muscle-specific desmin intermediate filaments. Furthermore, gene expression is similar to adult cardiomyocytes as cardiac-specific genes are expressed. HL-1 cells exhibit spontaneous

action potentials and synchronous beating in confluent cultures. The kinetics of the polarization are characteristic of cardiac myocytes and therefore HL-1 cells have been widely used as model systems to investigate the electrophysiological properties of cardiomyocytes[236]. Furthermore HL-1 cells have been implemented to characterize apoptosis, cell cycle, oxidative stress, signal transduction, transcriptional regulation in cardiomyocytes. They have also been used as model systems to elucidate the effects of common pathophysiological conditions such as hypoxia, hyperglycemia and hyperinsulinemia.

HL1 cells were provided by Prof. William C. Claycomb (Departments of Biochemistry and Molecular Biology and Cell Biology and Anatomy, Louisiana State University Medical Center, New Orleans, LA 70112) and cultured as described[237]. HL-1 cells were harvested for experiments at their maximum contraction.

3.1.2 Protein Analysis

3.1.2.1 Western Blot

Specificity of antibodies directed against histone modifications was confirmed by western-blots using whole calf thymus histones. 0.5 µg calf thymus histones per lane were separated on SDS-PAGE gels, transferred onto PVDF membrane and reacted with primary antibodies diluted as follows: rabbit anti-dimethylated H3K4 serum (1/250), rabbit anti-trimethylated H3K4 antibody (1/1000), rabbit anti-H3K9acK14ac Antibody (1/1000), anti-H4K5acK8acK12acK16ac (1/125). Incubation with first AB o/n in 0.1% Tween TBS at 4°C. The primary antibody reactions were detected with anti-rabbit IgG conjugated with HRP (1/10,000), followed by ECL Advance detection.

Specificity of antibodies directed against transcription factors was tested using HL-1 whole cell lysate. Whole cell lysate was prepared by incubating approximately 1×10^7 cells with 1ml of RIPA buffer supplemented with 20 µl of Roche Complete Protease Inhibitor and 1 µl of Benzonase for 30min on ice. Protein concentrations were determined with Bradford[238] analysis using the protein assay developed by Biorad according to the manufacturers instructions. 100 µg of whole cell lysate were separated 10% SDS-PAGE gels and transferred onto Nitrocellulose membranes and reacted with primary antibodies as follows: Gata4 sc-1237 antibody (1/200), Nkx2.5 sc-14033X antibody (1/50), Mef2a sc-313 antibody (1/50) and Srf sc-335 antibody (1/50). Incubation with first AB was carried out o/n at 4°C. In case of Gata-4 the primary antibody was diluted in 3% milk powder in PBS-T, in case of Nkx2.5, Mef2a and Srf primary antibodies were diluted in Crossdown Buffer (Applichem). The primary antibody reactions were detected with anti-rabbit IgG or anti-goat IgG as described above.

3.1.2.2 Indirect Immunofluorescence

siRNA treated or untreated HL-1 cardiomyocytes were fixed in 100% methanol at -20 °C for 10 min, permeabilized with 0.1% Triton X-100-phosphate-buffered saline (PBS) for 10 min at RT, washed twice with PBS 0.05% Tween and blocked with 3% bovine serum albumin-PBS/0.05% Tween for 30 min at r.t.

All primary rabbit antibodies were used at 1:200 dilution and mouse anti-αTubulin was used at 1:800 at 4 °C o/n. The cells were then washed 4x PBS 0.05% Tween and incubated with a fluorescent secondary red Alexa 568 anti-rabbit antibody at 1:500 dilution and green Alexa 488 anti-mouse antibody at 1:2000 for 45 min at r.t. Subsequently, cells were washed 4x with PBS 0.05% Tween, rinsed with ddH2O, and the cover slips were mounted on glass slides with Vectashield Mounting medium containing DAPI for nuclear staining.

Fluorescence was visualized using a Zeiss Axioimager.Z1 inverted fluorescence microscope equipped with x10 and x20 objectives. Images were acquired using Axiovision imaging software.

3.1.3 Nucleic Acid Analysis

3.1.3.1 Quantitative Real-Time PCR

Quantitative real-time PCR (qPCR) is a method widely used for the absolute or relative quantification of gene expression but can also be used to quantify relative amounts of any dsDNA as in ChIP analysis. Quantification is based on the measurement of fluorescence of the dye SYBR green I whose fluorescence increases ≈ 200 fold when intercalated into double stranded DNA. The intensity of the fluorescence is directly proportional to the amount of dsDNA in the PCR reaction.

Primers were designed using PrimerExpress software to amplify 100-150 bp fragments. Primers for verification of ChIP array experiments were designed to amplify genomic DNA regions with probes showing enrichment in case of positive controls or no enrichment in case of negative controls on the array data. Primers for verification of expression array data were designed to be exon spanning in order to avoid falsification of results in case of genomic DNA contamination. All used primers show linear amplification behavior as tested by standard curves and no detectible reaction products in no template control reactions. Amplification efficiency was calculated according to Swillens *et al.*[239] and was found to be comparable for all primers.

All qPCRs were measured on ABI Prism 7700 in 10 µl reaction volume with 2 times SYBR green I master mix and 100 nM primer in duplicate. Standard curves for primers

designed for ChIP experiments were measured on genomic DNA with 0.1 µg, 1 µg, 10 µg and 100 µg per well, for test of primers designed for expression analysis a dilution series of cDNA with 0.375 ng, 1.5 ng, 6 ng, 24 ng and 96 ng per well were used. Ct values were determined using the integrated SDS 2.1 software. Fold changes were calculated using the relative quantification method of ΔΔCt as described in the manufacturers manual (http://www.appliedbiosystems.com/support/tutorials/7700amp/). Fold changes for expression analysis were normalized to Hprt1. The scale of absolute expression levels as measured by real-time PCR was adjusted to the scale of the array intensities. Fold change enrichments of ChIP samples were measured relative to input.

3.1.3.2 Bioanalyzer Analysis

Bioanalyzer measurements are used to determine the quantity and quality of RNA, DNA or protein samples. The system is an gel electrophoresis based microfluidics system. The DNA or RNA sample is loaded onto a gel matrix in which the fragments are separated according to size. As in standard agarose gel electrophoresis a dye intercalates into the nucleic acids and the fluorescence of the dye changes proportionally to the concentration of DNA or RNA. The Bioanalyzer measures the fluorescence of the dye (proportional to DNA concentration) and plots it against the migration time (proportional to DNA size).

Due to the omnipresence of RNases Bioanalyzer measurements are of particular value to determine whether RNA is partially degraded before running microarray or qPCR applications. Furthermore the 2100 bioanalyzer software calculates an estimate of the quantity and ribosomal ratios of the total RNA sample. In case of DNA as obtained from ChIP experiments degradation can be detected and the distribution of the fragment sizes can be estimated. Measurements of DNA and RNA samples were performed according to the manufacturers instructions using the reagent kit DNA 7500 assay and RNA 6000 Nano assay, respectively. An example for the results obtained for the total RNA preparations is shown in Figure 3-1. The isolated total RNA samples showed no sign of degradation.

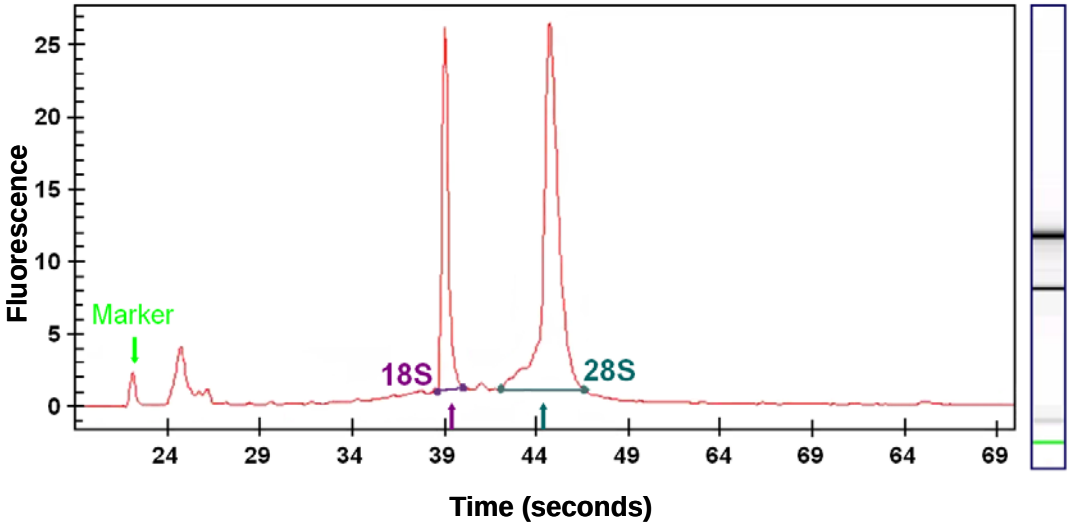

Figure 3-1. RNA quality analysis by Agilent bioanalyzer, exemplary for total RNA from C2C12 differentiated cells of the fourth passage. The first peak is a 20 bp molecular marker. The second peak consist of tRNA of approximately 80 bp and small RNAs[240]. The third and fourth peak are 18S and 28S rRNA, respectively.

3.1.3.3 Transfections

Transfection is the introduction of foreign material into eukaryotic cells. This is frequently achieved by mixing the material to be transfected, in this case the siRNA, with a transfection reagent (e.g. Lipofectamine 2000, a cationic lipid) to produce a liposome. The liposome fuses with the cell plasma membrane thereby depositing the cargo inside. For each cell line and often also for each oligonucleotide, transfection protocols have to be optimized to achieve a maximum amount of oligonucleotide in the cells while retaining minimum toxicity. Optimal conditions for transfection of HL-1 cells were determined and communicated by Dr. Christina Grimm, Max Planck Institute for Molecular Genetics Berlin, Germany.

siRNA Transfection into HL-1 Cells

HL-1 cells were cultured as described above, but all steps were carried out without addition of antibiotics in the cell culture media. Cells were grown for at least two days to 70-80% confluence. HL-1 cells were seeded into 6-well plates with 2 ml media containing 3×10^5 cells in each well, resulting in 70-80% confluence after settling for 4 h. 9 µl of 20 µm siRNA was mixed with 270 µl of DMEM (mix A) and 16 µl of Lipofectamine 2000 was combined with 470 µl DMEM (mix B). Mix A and mix B were combined within 5 min of preparation, incubated for 20 min at r.t. and the mixture was added drop wise to the cells. After 24 h the cell culture media was changed and after a further 24 h the cells were harvested and RNA was isolated as described in 3.1.4

3.1.4 Expression Array Hybridizations

3.1.4.1 Expression Analysis on NimbleGen Arrays

For each cell type (myoblasts, myotubes and cardiomyocytes) six isolations of total RNA from three different passages were performed using Trizol and subsequent DNase digest according to the manufacturers instructions. RNA from three independent isolations was pooled giving two biological replicates per cell type. RNA quality was confirmed by Bioanalyzer or agarose gel analysis and subsequently labeled and hybridized by NimbleGen standard procedures or using the Ambion Illumina labeling kit according to the manufacturers instructions.

To confirm array data cDNA was synthesized by reverse transcription using AMV-RT with random hexamers. 1 µg total RNA was denatured with 4 µl 5 mM dNTPs, 1 µl 1.2 µg/µl $pd(N)_6$ in a final volume of 10 µl DEPC treated water for 10 min at 65 °C and chilled on ice for 5 min. Subsequently 4 µl 5x AMV-RT buffer, 4 µl 25 mM MgCl2, 0.5 µl 40 U/µl Recombinant RNasin ribonuclease inhibitor and 1.5 µl 10 U/µl AMV-RT were added to a final volume of 20 µl. The reaction was carried out for 1 h at 42 °C. cDNA corresponding to 6.25 ng of initial total RNA was used as template for qPCR reactions. Expression array data were confirmed for 15 transcripts in each cell type by qPCR.

3.1.4.2 Expression Analysis on Illumina Arrays

For analysis of the function of transcription factors Illumina bead studio Mouse-6 v1.1 arrays were used. These arrays were not available at the time of the project investigating the histone modifications but have the advantage of covering the entire mouse genome. To ensure array quality first an aliquot of the HL-1 total RNAs that had been previously analyzed on the NimbleGen expression arrays were reanalyzed. The two platforms (NimbleGen and Illumina) were found to be in excellent agreement.

Labeling of the total RNA samples derived from HL-1 cells as well as siRNA treated HL-1 cells was carried out using the AMIL1791 Illumina® TotalPrep RNA Amplification Kit (Ambion) according to the manufacturers instructions. Labeled RNA was hybridized and the intensities were scanned by Integragen (France).

3.1.5 Chromatin Immunoprecipitation (ChIP)

3.1.5.1 ChIP Histone Modifications

ChIP experiments for were performed in duplicate for five different antibodies in parallel with modifications as described[241]. Briefly $\approx 10^8$ cells (C2C12 undifferentiated, differentiated or HL-1) were grown as described in the section cell culture (0 and 0).

Formaldehyde was added directly to culture medium to a final concentration of 1% and cells were incubated for 10 min at 37 °C. Subsequently cross-linking was quenched by adding glycine to a final concentration of 125 mM. Cells were washed twice with 4 °C phosphate-buffered saline, collected and sedimented at 450 x g for 10 min at 4 °C. Cells were swelled for 10 min on ice in hypotonic buffer, collected by centrifugation, resuspended in hypotonic buffer and lysed with a Dounce homogenizer. Nuclei were collected by centrifugation for 15 min at 20,000 g at 4 °C and resuspended in sonication buffer. The chromatin was fragmented by sonication with a Branson 450 sonifier to an average size of 600 bp and cell-debris removed by centrifugation. For immunoprecipitation buffer conditions were adjusted to RIPA conditions by adding RIPA concentrate buffer.

Chromatin was aliquoted to five separate samples for immunoprecipitation and a fraction of material was saved as 'chromatin input'. Chromatin was precleared by rotation with Protein A/G beads for 1 h at 4 °C. To each immunoprecipitation one of the following polyclonal antibodies was added and rotation continued over night: Anti-acetyl-Histone H3 (10 µl, 1 ng/µl), Anti-acetyl-Histone H4 Antibody (10 µl; 1 ng/µl), Histone H3 (dimethyl K4) antibody (4 µl, 200 ng/µl), Histone H3 (trimethyl K4) antibody (4 µl, 400 ng/µl) and Rabbit Normal IgG (5 µl, 400 ng/µl). Protein A/G beads were added and rotation continued for 1 h.

Immune complexes were washed five times at 4 °C for 10 min each with following buffers: twice with RIPA buffer, RIPA buffer with 500 mM NaCl, Li/Detergent solution and TBS. Immunocomplexes were disrupted by first eluting 10 min at 65 °C with 1% SDS/TE buffer and a second elution for 15 min with 0.67% SDS/TE buffer. Eluates were pooled and cross-links disrupted by heating at 65 °C over night. Subsequently DNA was treated with RNase A, Proteinase K, purified by extraction with phenol-chloroform/isoamylalcohol and chloroform and finally ethanol precipitated.

To validate the quality of ChIP experiments for histone modifications before linear amplification, qPCR was performed for 3 histone modified sites and 4 non-modified sites per cell-line and modification as described above. Normal rabbit ChIPs gave no enrichment over input for any of these sites and yielded less than 1% DNA compared to specific antibodies and therefore did not yield enough DNA to amplify for 'on chip' applications.

Amplification of ChIP and Input DNA

Linear amplification of ChIPed DNA and input control was carried out on the basis of random primer amplification developed by Bohlander *et al.*[202] and which was subsequently modified for ChIP applications[242] except only one round of amplification with 20 cycles was performed. Amplified samples were purified using Wizard SV PCR purification kits

according to the manufacturers instructions. DNA quality was confirmed by Bioanalyzer measurements. Samples were labeled and hybridized according to NimbleGen standard procedure.

After bioinformatics analysis of the ChIP-chip data, a subset of modification sites was validated by real-time PCR for locations of 9 non-modified sites and 12 modified sites per modification and cell type as described above (in total 252 single verifications).

3.1.5.2 ChIP-chip against Gata4, Mef2a, Nkx2.5, and Srf

Chromatin immunoprecipitation experiments directed against the TFs Gata4, Mef2a, Nkx2.5, and Srf were carried out in HL-1 cells as described for histone modifications. The following antibodies were used at concentrations of 2 µg/ml: Gata4 sc-1237 antibody, Nkx2.5 sc-14033X antibody, Mef2a sc-313 antibody, and Srf sc-335 antibody. As controls Rabbit Normal IgG and Goat Normal IgG were used at the same concentrations. To obtain enough material for hybridization, each ChIP experiment was performed multiple times and independently amplified for 20 or 22 cycles in reaction B. The amplified ChIPed material and Input was combined from between two and four experiments resulting in two independent pools for each TF. The enrichment of known target genes was confirmed in each separate experiment (data not shown) and in the two independent pools (Chapter 10.2) As for the ChIP experiments directed against histone modifications near to no DNA was precipitated using the Rabbit Normal or Goat Normal IgGs.

3.2 Design of Arrays

Arrays Used for Analysis of Histone Modifications: Expression Array and Histone ChIP-Array

Human or mouse transcripts expressed in heart, skeletal or smooth muscle were selected from several sources as listed in Table 3-1. The final lists comprised 12,625 unique mouse transcripts. All identifiers were mapped to Ensembl version 26, human-mouse orthologs were identified and redundant entries were removed. These transcripts were selected to be represented on the expression arrays. For the respective set of genes, the human-mouse conserved non-coding blocks (CNBs) in the 5 kb region upstream of annotated TSSs and in the first intron up to 10 kb downstream of each TSS were considered. For genes with less than 10% CNB sequence, a fixed region of 2.2 kb upstream and 0.8 kb of the first intron was selected. Additionally, the first exon of each gene was represented. The selected regions were repeat-masked and probes were designed by NimbleGen. The resultant array

design represents 8,585 genes of the mouse genome with 390,000 probes of 70 bp length and 85 bp between probes.

Table 3-1. Sources of transcripts represented on expression and ChIP arrays.

Source	Number of transcripts
Key genes of cardiac development	55
Human chromosome 21 transcripts in Ensembl v26	211
Manually selected controls	204
Transcripts expressed in human heart Kaynak *et al.*[243]	2,546
Symatlas human atrioventricularnode – A/B [244]	2,399 / 2,399
Symatlas human cardiac myocytes – A/B[244]	4,786 / 3,981
Symatlas human heart – A/B[244]	3,391 / 3,978
Symatlas human skeletal muscle – A/B[244]	1,889 / 1,761
Symatlas human smooth muscle – A/B[244]	5,296 / 5,237
Symatlas mouse heart[244]	1,665
Symatlas mouse skeletal muscle[244]	1,793
Transcripts expressed in mouse hearts Tabibiazar *et al.*[245]	132
All transcription factors listed in Transfac[246] as of Jan 2005	2,236

Arrays Used for Analysis of ChIP Experiments Directed against Transcription Factors: TF ChIP-Array

For the second generation of arrays for ChIP-chip analysis the same set of genes was represented as before (Table 3-1). However, the annotation of Ensembl mm8 v39 was used. Furthermore, for each transcript 2 kb upstream and 100 bp downstream of the annotated TSS was represented. Additionally, the conserved non-coding blocks (CNBs) in the 10 kb region upstream and 3 kb downstream of annotated TSSs were considered. Bases were considered to be conserved if annotated with a Phastcons value[247] of at least 0.2. Conserved regions were merged if less than 300 bp apart and enlarged to a minimum size of at least 1 kb. For the selected regions containing approximately 89 Mbp probes were designed by NimbleGen without masking of repetitive regions. The probes were then compared to the mouse genome build mm8 and probes with multiple hits in the genome were removed. The final array design contains 740,000 probes with approximately 50-60 bp probes and a tiling of 110 bp (50-60 bp gap between probes).

3.3 Data Analysis

3.3.1 Histone Modification Array Analysis

Histone ChIP Microarray Preprocessing

Probes were mapped to the mouse genome assembly *mm8* using BLAT [248], allowing up to one mismatch per 50 mer probe, resulting in 389,918 genomic positions. Intensities of each channel were normalized and log-transformed using VSN [249]. Log-ratio enrichment levels for each probe were calculated by subtraction of log Cy3 (input) from log Cy5 (ChIP sample).

Identification of Histone Modified Sites

Normalized probe levels for biological replicates were averaged and smoothed along chromosomal coordinates using a sliding window method. For each probe position the smoothed probe level was computed as the median over the probe levels in a 800 bp window centered at that position. To allow for different efficiencies of antibodies, a cut-off was defined for each type of histone modification separately, by repeating the above smoothing procedure on data where probe positions were randomly permuted and calculating the 99% quantile. A probe was called enriched if it had a smoothed probe level greater than the cut-off. Enriched probes were merged into enriched regions if less than 600 bp apart. Resultant regions of at least three probes were called *modified sites*.

Identification of Histone Modified Domains

A *modified domain* is a combination of different modified sites within one cell type. The genomic locations were required to overlap by at least 75% of the length of the smallest contained modified site. A domain was mapped to a TSS if its middle position was located up to 5 kb upstream or within the respective transcribed region.

NimbleGen Expression Microarray Preprocessing

Probes were mapped to the mouse genome assembly *mm8* using BLAT [248], allowing up to one mismatch per 24mer probe. Only probes matching annotated Ensembl transcripts (version 39, June 2006) were further analyzed. Probe intensities were background-corrected, quantile normalized and summarized into transcript expression levels using the median-polish procedure [250]. Analysis of microarray data showed that arrays representing biological replicates have high correlation coefficients of 0.99 (Figure 3-2 A), indicating excellent reproducibility of data and very good array quality. Cross-cell comparisons (Figure 3-2 B)

give lower correlations indicating a high number of differentially expressed genes as would be expected from different cell lines.

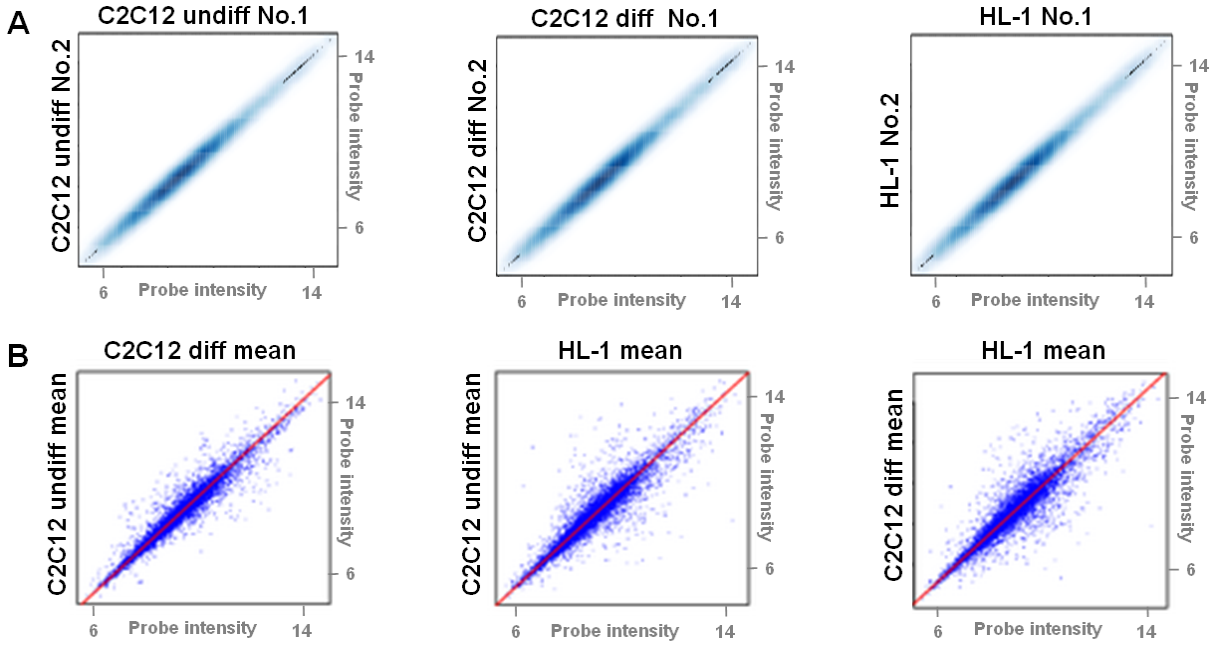

Figure 3-2. Correlation of expression arrays. A) Correlation between biological duplicates from one cell type is 0.99. B) Duplicates from one cell type are averaged and correlation between cell types is shown. Correlation coefficients range between 0.93 in case of comparison between HL-1 and C2C12 differentiated cells and 0.96 in case of undifferentiated versus differentiated C2C12 cells.

Expression Level Differences between Transcript Categories

Transcripts were categorized according to modified sites or modified domains. Pairwise Wilcoxon rank sum tests were employed to assess differences in expression between transcript categories. The resulting *p* values were corrected for multiple testing using the Bonferroni procedure[251]. A corrected $p \leq 0.05$ was interpreted as evidence of transcripts in one category having significantly different expression levels from those in the other category.

Transcript Expression Categories

Transcripts were categorized according to expression levels as non- (<7 arbitrary scanner units, $\log_2$ scale), low- (7-8.5) medium- (8.5-10.9) and highly- expressed (>10.9). The first three cut-offs were determined by RT-PCR for transcripts of known transcriptional status. The cut-off for high expression is the 90% quantile of all data above 8.5.

Identification of Differentially Expressed Genes

Differential expression of transcripts between cell types was assessed using the moderated t-statistic of the empirical Bayes approach in the Bioconductor[252] package limma[253]: *p* values for differential expression were adjusted for multiple testing using

Benjamini's and Yekutieli's method for control of the false discovery rate[254]. Transcripts with an adjusted $p \leq 0.05$ were considered to be differentially expressed.

Linear Model: Expression Levels

For the regression of the normalized transcript expression levels on histone modifications, the data of undifferentiated and differentiated C2C12 cells was used. A linear regression model was fitted to include for each transcript: expression levels in each cell type, presence of modified sites and median GC content of the microarray probe that mapped to the transcript (Model 3-1, all models are given in S-plus/R formula notation). A *t*-test was computed for the coefficient of each effect to assess whether it was significantly different from 0 (Table 4-6). *p* values were corrected for multiple testing using the Bonferroni procedure[251]. Certain interaction terms between modifications were significantly different from zero. This confirmed the non-additivity of the modification effects.

Model 3-1. Linear Model: Expression Levels. Given is a linear model relating absolute expression level to cell type, presence of modified sites, median probe GC content and interactions.

y ~ H3ac + H4ac + H3K4me2 + H3K4me3 + GC + H3ac:H4ac + H4ac:H3K4me2 + H4ac:H3K4me3 + H4ac:H3K4me2:H3K4me3 + H3ac:H4ac:H3K4me2 + H3ac:H4ac:H3K4me3 + H3ac:H4ac:H3K4me2:H3K4me3 + H3ac:H3K4me2:H3K4me3 +H3ac:H3K4me2 + H3ac:H3K4me3 + H3K4me2:H3K4me3 + cell.type

where
y: expression level of transcript in cell line
H3ac: indicator variable for transcript's associated modification H3ac; it is 1 if at least one H3ac is associated to the transcript, 0 otherwise.
H4ac, H3K4me2, H3K4me3: analogous to H3ac
GC : median percent GC content of all expression microarray probes mapped to transcript
cell.type : one of “C2C12U”, “C2C12D” or “HL1”
The expression “A:B” denotes the interaction term between predictors A and B.

Logistic Regression Model

For the logistic regression, the response was an indicator variable for differential upregulation or downregulation of a transcript, while the predictors were indicator variables for transcripts gaining or losing modifications during differentiation of C2C12. Model 3-2 is for genes gaining modifications, the coefficients are given in Table 4-7. Model 3-3 is for genes losing modifications, the coefficients are given in Table 4-8. The data of undifferentiated and differentiated C2C12 cells was used. *p*-values were corrected for multiple testing using the Bonferroni procedure[251]. Gain of H4ac was significantly associated with differentially upregulated genes. 62 out of 126 up-regulated transcripts, however, did not show any modification gains.

Model 3-2. Logistic regression model: Associating modification gains during differentiation to differentially up-regulated genes.

Logistic regression analysis of binary indicator variable for upregulation between undifferentiated and differentiated C2C12 cells against gain of modified sites.

dy ~ dH3ac + dH4ac + dH3K4me2 + dH3K4me3

where
dy: indicator variable: 1 if the transcript is found at significantly higher level in differentiated C2C12 cells than in undifferentiated cells, 0 otherwise
dH3ac: factor variable for transcript's H3ac modification change with two levels: gain or no change. For gain, the transcript had no H3ac modification in undifferentiated cells but in differentiated cells.
dH4ac, dH3K4me2, dH3K4me3: analogous to *dH3ac*
Because individual observations are actually matched pairs of transcripts before and after differentiation and the median probe GC content stays constant during differentiation, we did not include it as a predictor in this model.

Model 3-3. Logistic regression model associating modification losses to differentially down-regulated genes.

Logistic regression analysis of binary indicator variable for downregulation between undifferentiated and differentiated C2C12 cells against loss of modified sites.

dy ~ dH3ac + dH4ac + dH3K4me2 + dH3K4me3

where
dy: indicator variable: 1 if the transcript is found at significantly lower level in differentiated C2C12 cells than in undifferentiated cells, 0 otherwise
dH3ac: factor variable for transcript's H3ac modification change with two levels: loss or no change. For loss, the transcript had a H3ac modification in undifferentiated cells but not in differentiated cells.
dH4ac, dH3K4me2, dH3K4me3: analogous to *dH3ac*

Gene Ontology Associations to Gene Groups

To analyze the association of differentially expressed transcripts with GO categories, the transcripts were mapped to genes. The association of gene groups to Gene Ontology[255] (GO) terms was assessed according to Alexa *et al*.[256] through a conditional hypergeometric test for overrepresentation using a *p* value threshold of 0.001. In a pre-filtering step, we excluded from this analysis all transcripts whose expression levels across samples had an inter-quartile range below 0.5 $\log_2$ units.

Implementation

The computational methods were implemented in the R programming language, using packages from the Bioconductor project [252] and extending these with custom source code. An R package containing functions used in the analysis has been published by Toedling *et al*.[257].

Mef2 Binding Site Analysis

The occurrence of the TRANSFAC[246] Mef2 binding site V$MMEF2 was compared in a 1.5 kb window between 816 domains with H4ac-gain and 790 domains with H4ac-loss during differentiation. MAST[258] was used with a minimum sequence *p* value of 0.1 and an

E value of 160. The Group Specificity Score[259] was used to identify significant overrepresentation.

3.3.2 TF-ChIP-Array Analysis

Identification of TFBSs

Intensities of each channel were normalized and log-transformed using VSN [249]. Log-ratio enrichment levels for each probe were calculated by subtraction of log Cy3 (input) from log Cy5 (ChIP sample). Signals were smoothed by calculating a median over the probes inside a sliding window of size 600 bp. To distinguish enriched probes a z-score and empirical p value for each probe on the null hypothesis that these z-scores have a symmetric distribution with mean zero was calculated. P values were corrected for multiple testing and probes with a nominal false discovery rate of smaller 0.1 were considered to be significantly enriched. Significant probe positions having less than 210 bp between each other were combined into peaks by single-linkage clustering and considered as true TF binding sites. Identified TFBS were assigned to the 12,625 represented TSSs if located within 10 kb upstream or 3 kb downstream.

Comparison of TFBS Sequences to Annotated TFBM

For each TFBS of a particular TF the sequence surrounding ±250 bp of the peak center was analyzed for the occurrence of known binding motives for that particular TF. For this purpose all annotated TFBM for the respective TFs (Table 3-1) were extracted from Transfac[260] and matched to the binding site sequences using the Transfac MATCH™ program[261]. It was analyzed how many TFBS contain TFBM and how often TFBM occur multiple times in one TFBS. Furthermore, it was investigated how often TFBM lie within conserved regions, several different conservation criteria were applied and compared

Table 3-2. List of Transfac identifiers of TFBM used.

Gata4	Mef2	Nkx2.5	Srf
V$GATA4_Q3	V$MEF2_01	V$NKX25_01	V$SRF_01
V$GATA_Q6	V$MEF2_02	V$NKX25_Q5	V$SRF_Q6
	V$MEF2_03		V$SRF_C
	V$MEF2_04		V$SRF_Q4
	V$AMEF2_Q6		V$SRF_Q5_01
	V$MMEF2_Q6		V$SRF_Q5_02
	V$HMEF2_Q6		
	V$MEF2_Q6_01		

Conservation Analysis

To analyze the degree of conservation of TFBS different conservation criteria were defined. PhastCons Conserved Elements[247] are a prediction of conserved elements based on *PH*ylogenetic *A*nalysis with *S*pace/*T*ime models of five vertebrate species (human, mouse, rat, chicken, and *Fugu rubripes*), four insect species (three species of *Drosophila* and *Anopheles gambiae*), two species of *Caenorhabditis*, and seven species of *Saccharomyces*. Additionally a 70% conservation between human-mouse was defined (MH70), where a base is called conserved if at least one 100 bp window can be found that includes this base and where at least 70% of the bases in this window are conserved. A 100% conservation between human and mouse (MH100) was defined where only such bases are considered with exact conservation. Additional analyses were performed without any conservation masking.

***De Novo* Motif Search**

A *de novo* motif search was performed using the sequence ±250 bp of the peak centers. These sequences were masked using different conservation criteria and the following motif search programs were used: Bioprospector[262], AlignACE[263], MEME[264] and Weeder[265]. Retrieved motifs were compared to TFBM as annotated in Transfac (Table 3-2).

Occurrence of TFBSs

It was analyzed how often multiple TFBSs were assigned to the same transcript irrespective of the distance between the TFBS (co-regulation analysis). In a second approach it was investigated how often TFBS occur within a 500 bp window (co-binding analysis). Information on the position of CpG islands was extracted from UCSC Genome Browser[266] and a TFBS was said to lie within a CpG island if the TFBS-center was located within ± 500 bp.

Position of TFBSs Relative to Histone Modifications

As the Histone-ChIP array covers significantly less of the genome compared to the TF-ChIP array, in this analysis only such TFBS were considered that are located within regions also represented on the Histone-ChIP array. A TFBS was considered to lie within histone modified sites or regions if the region defined by the TFBS-center ± 500 bp lies inside the histone region or partially overlaps with it.

Illumina Array Expression Analysis

First probes are filtered according to the detection score given by the Illumina array analysis software BeadStudio (Illumina proprietary software). Only probes with a detection

score ≥ 0.95 are retained. Probe intensities are qspline normalized and intensities mapping to one transcript (Ensembl mm8 v46) are summarized using the median polish procedure. Differential expression was determined using the limma package[253]: *p*-values were corrected for multiple testing according to Benjamin and Yekutieli[254]. Transcripts with $p \leq 0.05$ were considered to be significantly differentially expressed.

3.4 Figures and Graphs

Photographs were adjusted for brightness and contrast using Microsoft PowerPoint. Schematic Figures in section 1 and the graph in Figure 4-25 were drawn using Microsoft PowerPoint and ScienceSlides (http://visiscience.com/). Figure 5-3 was constructed using BioTapestry[267], an open source software package available from http://www.biotapestry.org. Computation of the graph layout in Figure 4-18 was realized using the Rgraphviz[268] package from R-Bioconductor.