Bioinformatic Analysis of Cardiac Transcription Networks

Markus Schüler

April 2011

DISSERTATION

zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Silke R. Sperling

1. Gutachter: Prof. Dr. Martin Vingron

2. Gutachter: Prof. Dr. Silke R. Sperling

Disputation: 8. Juli 2011

for my family

Contents

Danksagung

Diese am Max-Planck-Institut für molekulare Genetik durchgeführte Studie wurde gefördert durch ein Stipendium der Max-Planck-Gesellschaft und das EU Forschungsprojekt „HeartRepair“. Im Folgenden möchte ich mich herzlich bei all denen bedanken, die mich bei der Durchführung unterstützt haben und durch die es eine schöne Zeit geworden ist:

Prof. Dr. Silke R. Sperling dafür, dass sie das alles ermöglicht hat, für ihre intensive Betreuung, ihren Ideenreichtum und für das offene Ohr auch abseits aller wissenschaftlichen Fragestellungen. Zudem für die Möglichkeit, auf vielen Konferenzen zu Gast gewesen sein zu dürfen.

Prof. Dr. Martin Vingron für die bioinformatische Erdung dieser Arbeit, seine Supervision als Doktorvater und seinen Einsatz für die Belange der Studenten am MPI.

Allen derzeitigen und früheren Mitarbeitern der AG Sperling:
Marcel Grunert, für unsere mal mehr mal weniger wissenschaftlichen Diskussionen, die gemeinsame Zeit auf vielen Konferenzen und seine ansteckende gute Laune. Jenny Schlesinger für ungezählte Apfelstückchen, ihr Engagement im Labor und ihren unerschütterlichen Unglauben an bioinformatische Analysen. Beiden danke ich auch für die umsichtige Korrektur dieser Doktorarbeit. Cornelia Dorn für das freundlichste Lächeln der Arbeitsgruppe, den „Conny“ und die netten Unterhaltungen beim Warten auf das Teewasser. Ilona Dunkel für ihre einmalig direkte Art, die viele Arbeit im Labor, unsere Diskussionen über soziales (Fehl-)Verhalten und die Tatsache, daß sie mir endlich glaubt, dass ich meine Doktorarbeit gemacht habe. Qin Zhang für den Einsatz im Labor und die Einführung in echte chinesische Esskultur. Barbara Gibas für all den abgenommenen Verwaltungsaufwand, die alternativen Meinungen und für die ständige Versorgung mit Dickmachern. Martje Tönjes, weil sie die netteste Top-Terroristin der Welt ist, unser gemeinsames Weihnachts-EXPA-Paper-Schreiben, viele zu analysierende Daten und die Gewißheit, dass es noch echt gute Menschen auf dieser Welt gibt. Tammo Krüger dafür, dass er mir in seiner ganz eigenen Art gezeigt hat, was wirklich wichtig ist, und dass er der Einzige ist, der wirklich Ahnung von guten Filmen hat. Jenny J. Fischer für die vielen Experimente und weil sie mir lehrreiche Einblicke in die menschliche Psyche gewährt hat. Christina Grimm, Martin Lange und Alan Punnoose für ihre Arbeit im Labor und viele nette Stunden. Stefanie Hammer und Siegrun Mebus für die viele Arbeit mit den Patientendaten.

Zudem all den Mitarbeitern des MPI, die mir bei wissenschaftlichen Fragen weitergeholfen haben, darunter Szymon Kielbasa, Thomas Manke, Utz J. Pape und Helge Roider. I like to thank Paz Polak for beeing special in his own way, our time together and our co-work for the STA. Akdes Serin and Christian Roedelsperger for so many nice chats. Cornelia Lange für unser Teamwork in der STA.

Allen CHD-Patienten und ihren Familienangehörigen, die an dieser Untersuchung teilgenommen haben, zudem allen Mitarbeitern und Co-Autoren auf unseren Papern, ohne die das alles nicht möglich gewesen wäre.

Meinen Freunden für all die Zeit, die ich nicht mit Wissenschaft verbracht habe, und für die Realisierung von „Projekt Nagoya“.

Meinen Eltern und Großeltern, die mir das alles ermöglicht haben: Meinem Vater und Angelika für eure Ermutigungen. Meiner Mutter für all die Liebe, die sie mir so uneingeschränkt schenkt.

Christophe Mendes, weil er die ganze Zeit an meiner Seite war und ist und weil er mir immer wieder zeigt, wo die wirklich großen Schätze des Leben zu finden sind.

Thesis Outline

A panel of key transcription factors are the main drivers of the cardiac developmental process and are essential for normal cardiac function. However, limited insights have been generated at a systems-level about how these factors modulate the overall transcription network, how they act in a combinatorial manner and moreover how they interplay with epigenetic or environmental factors. To tackle these open points, a systems biology approach was chosen such that well-defined high-throughput experiments were used as a starting point and based on their outcome subsequent experiments followed. Gene-focused hypothesis driven laboratory experiments were performed generating time-series data to puzzle down a sequence of transcription factor binding, histone modification and respective gene transcription. Finally, extracted transcription networks were studied and extended based on expression profile disturbances in diseased human hearts. This thesis represents the bioinformatics part of this overall project and aimed to provide the best suitable bioinformatic and statistical data analysis, predicted new transcription networks and proposed consequent laboratory experiments.

In the course of the study, five essential datasets were generated and analyzed: (a) genome-wide ChIP-chip data of the cardiac transcription factors Srf, Mef2a, Gata4 and Nkx2.5 in cardiac cell culture, (b) microarray gene expression profiles of wildtype and RNAi treated respective cell culture, (c) ChIP-seq data for Srf and histone 3 acetylation in cardiac cell culture, (d) time series ChIP-qPCR data of Srf, p300 and histone 3 acetylation and methylation in mouse hearts at E18.5, P0.5 and P4.5, and (e) gene expression profiles of human diseased hearts. Alongside, a panel of different bioinformatics and statistical methods suitable to analyze these datasets were identified. Their advantages and disadvantages are discussed and knowledge gained in the course of the project is presented. Finally, cardiac transcription networks were predicted based on the wealth of the data which could to a great extent be confirmed in respectively designed follow-up experiments. As a final step, the Cardiac Regulatory INteraction database (CARIN) was built to integrate data from this project with publicly available and relevant datasets as well as cross-species datasets obtained within the European FP6 project “HeartRepair”. The work presented in this thesis was published in two articles (Molecular BioSystems 2008 and PLoS Genetics 2011), one further manuscript is in preparation.

All analyses described have been performed in the group of Prof. Dr. Silke R. Sperling at the Max Planck Institute for Molecular Genetics, Berlin.

1. Introduction

1.1 Regulation of Eukaryotic Gene Transcription

The process of deciphering genotypic information into phenotypic characteristics is the main purpose of eukaryotic cells. Due to the central dogma of molecular biology, genetic information is stored in DNA, which is, after transcription into RNA, translated into proteins, which are the main functional units of every cell. Thus, a fine-tuned regulation of transcription plays an essential role in the maintenance of cellular function.

Transcription of genomic DNA into mRNA in eukaryotes is performed by RNA polymerase II, a complex made up of 12 different proteins. These are recruited to the transcription start site (TSS) by a combination of ubiquitously expressed proteins like the TATA-binding protein, which bind to the basal promoter, and tissue-specific transcription factors (TFs), that bind to factor specific *cis*-regulatory elements called transcription factor binding sites (TFBS) in the promoter regions of genes. Based on the binding of activating or repressing proteins, *cis*-acting elements are classified into enhancers (activator-binding) or silencers (repressor-binding). Just recently an additional group of *cis*-regulatory elements has been discovered. So-called insulators hinder activating proteins to propagate their function on to distant TSS. For a number of TFs their binding site preferences have been revealed, most commonly represented using position weights matrices (PWMs).

In turn, the ability of transcription factors to bind these *cis*-regulatory elements is highly dependent on their accessibility. DNA in the nucleus is highly condensed into a structure called chromatin by the use of basic proteins known as histones. Histones are complexes comprising two molecules of H2A, H2B, H3 and H4 each, as well as the linker histone H1. 1.75 turns of the DNA double-helix are tightly wrapped around the complex, building the nucleosomes.[1] A high compaction of the chromatin hinders the binding of TFs and therefore silences the transcriptional apparatus. Consequently, changes between the condensed (heterochromatin) and open (euchromatin) state are a key epigenetic factor for the regulation of gene expression. At least three distinct types of nucleosomal alteration have been proposed and proven to influence transcription levels of targeted genes: chromatin remodeling, core histone replacement and histone tail modifications.

An effective way of opening the chromatin structure is the removal or sliding of the histone octamer leading to nucleosome depleted regions of DNA. Interestingly, the same complex that forms chromatin structure during replication was found to be relevant for its remodeling.[2,3] Using ChIP-chip experiments in yeast it was shown that there is a positive correlation between the presence of nucleosome depleted regions located in the promoter of genes and the rate of gene transcription.[3] However, reports about corresponding nucleosome-free regions in the promoter of human genes are contradictory.[4,5] Further, bioinformatic studies have shown that *in vivo* nucleosome positioning is guided by the nucleotide sequence of wrapped DNA,[6-9] with nucleosome depleted regions mainly associated to poly-A/T stretches. In addition, replacement of histone particles by specialized variants has been shown to occur in the vicinity of transcribed regions (e.g. H2A.Z and H3.3[3,10]).

The best-studied nucleosome alterations are histone tail modifications. These post-translational modifications can influence the wrapping of DNA around the histone core and thereby lead to an altered transcriptional accessibility. More than 70 different histone modification sites are known, comprising acetylation, methylation, phosphorylation, ribosylation, sumoylation and ubiquitination[11] and a subset of these is shown in Figure 1. The influence of histone modifications on the accessibility

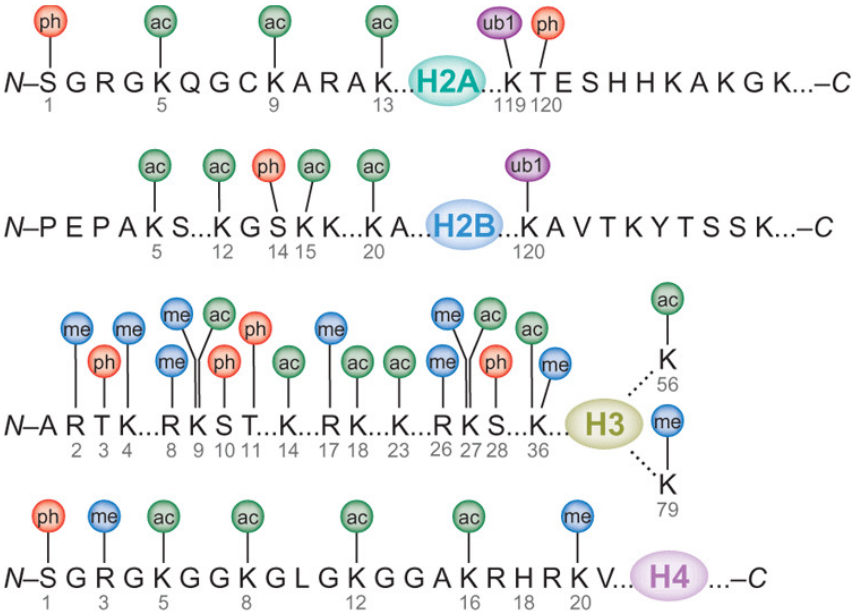

Figure 1: Histone tail modifications
Some of the known histone tail modifications including acetylation (ac), methylation (me), phosphorylation (ph) and ubiquitination (ub1). Figure taken from Bhaumik *et al.*[18]

of the wrapped DNA can either be direct ("direct effector model") or indirect through an effector protein/complex ("effector-mediated model"). An example for a direct effect is the neutralization of the positively charged histone tails through acetylation. These acetylation marks are set by a group of enzymes called histone acetyltransferases (HATs). This neutralization lowers the electrochemical coupling between the histone octamer and the wrapped DNA making the DNA more accessible for the transcriptional machinery.[12] Correspondingly, histone acetylation is tightly linked to an increase in transcription level.[13] Deacetylation, on the other hand, which is mediated by histone deacetylases (HDACs), is associated to a decreased expression level. In contrast, the well-studied SWI/SNF complex is an example for a recruited chromatin remodeling complex that actively alters chromatin structure through ATP hydrolysis representing the effector-mediated model.[14] These complexes might be recruited by DNA binding TFs or even through binding of modified histone tails. A possible candidate carrying the binding information might be histone tail methylation, which was shown to affect transcription level. In contrast to acetylation, methylation is not by itself associated to activated expression but seems to have a position-specific regulatory function. Methylation of the lysine at amino acid 4 of histone 3 (H3K4) was for instance shown to be associated to chromatin structures that allow transcriptional activation.[4] In contrast, methylation of lysine 9 at the same histone is thought to be linked to heterochromatin formation.[15] Furthermore, the degree of methylation of H3K4 was shown to be dependent on its genomic position. It decreases continuously from 5' to 3' with trimethylation at the 5' end, dimethylation in the middle and monomethylation at the 3' end.[2] The idea that distinct patterns of histone modifications can lead to transcriptional activation or repression different from the sum of their individual contributions, originally formulated by Strahl and Allis,[12] was named the "histone code hypothesis" and has since been studied in more detail.[16,17]

Another epigenetic regulatory mechanism, which is not analyzed in this study but which is of high importance for transcriptional regulation, is DNA methylation. Established by DNA methyltransferases, DNA methylation occurs in 60-90% of all mammal CpG sites,[19] however, recent studies found methylation also at non-CpG sites.[20,21] The main exceptions of the overall CpG

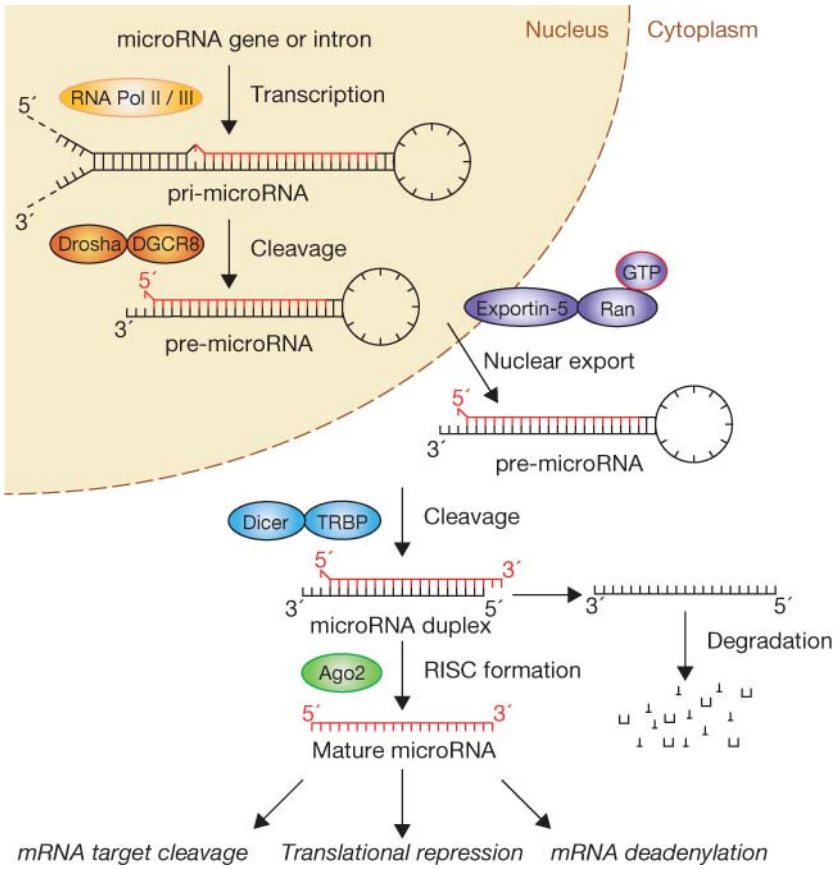

Figure 2: The miRNA processing pathway
MiRNA processing, transportation and their post-transcriptional regulatory functions. Figure taken from Winter *et al.*[23]

methylation are CpG islands in promoter regions, which appear to be mainly unmethylated. While CpG islands are associated with a large number of gene promoters, methylated CpG sites in the promoters of genes are associated to gene silencing, either through direct inhibition of TF binding or more likely by the recruitment of methylated DNA-binding proteins which in turn interact with additional proteins such as HDACs.[19] In this line, hypermethylation of DNA was found to be tightly coupled to heterochromatin formation and *vice versa*.[22]

In addition to the regulation of transcription, cellular abundance of RNA is regulated post-transcriptionally. RNA binding proteins regulate RNA splicing, RNA processing, nuclear export and nuclear degradation. RNA degradation or decay directly influences the amount of mRNA that can be transcribed into proteins and its regulation therefore enables a rapid alteration of protein synthesis. This area of study has recently gained more importance due to the increasing evidence that post-transcriptional regulation plays a crucial role in the regulation of gene expression. A recent study by Cheadle *et al.*[24] showed that during T-cell activation 55% of significant changes at the steady-state level had no corresponding changes at the transcriptional level, meaning they were the result of RNA stability regulation alone. A common way to regulate mRNA decay is the shortening of the poly-A tail by specialized exonucleases. This shortening is believed to destabilize the mRNA's association to cap binding complexes, which finally leads to its decay. Although RNA binding proteins may regulate post-transcriptionally large amount of the transcriptome, the targeting of a single gene is of special interest for the scientific community. Just recently, microRNAs (miRNAs), short RNAs with an average length of 22 nucleotides, were found to be important regulators of gene expression. MiRNAs

bind to mRNAs with complementary sequence in their 3' untranslated region (UTR), usually resulting in gene silencing.[25,26] In line with this, genome-wide computational and transcriptome analyses showed that the expression of miRNAs is more positively than negatively correlated with that of their targets.[27] The growing evidence for the high abundance of miRNAs in the cell further suggests that miRNAs may themselves be mediators of default repression.

MiRNAs are often transcribed from their own gene or are part of an mRNA gene's intron. They are transcribed into RNA hairpin loops by RNA polymerase II or III, capped to form pri-miRNAs and then spliced by the Drosha protein into pre-miRNAs (Figure 2). After export to the cytoplasm, pre-miRNAs are cleaved by the Dicer protein into the miRNA/miRNA* duplex. Although each strand of the duplex may potentially act as a functional miRNA, only one strand is usually incorporated into the RNA-induced silencing complex (RISC). The RISC, which contains a number of argonaute proteins, binds to the complimentary mRNA and cleaves it directly or recruits additional proteins to achieve translational repression or mRNA deadenylation followed by mRNA decay.[28] Current estimations predict over 1000 miRNAs in human,[29] targeting about 60% of all mammalian genes[30] thereby repressing hundreds of mRNAs each.[31,32]

The genetic, epigenetic and post-transcriptional levels of transcript regulation are closely related. Chromatin state directly influences the ability of TFs to bind their binding sites. In turn, the presence of certain TFs recruits histone modifying enzymes, which can lead to global and local changes in the chromatin state. MiRNAs primarily work at a post-transcriptional level. However, a recent study by Tan *et al.*[33] demonstrated that a miRNA can also directly influence the transcription of a gene by binding to its promoter, potentially coupled to the presence of a specific histone modification. In addition, a direct link between miRNAs and chromatin remodeling has been suggested.[34,35]

To summarize, eukaryotic gene transcription is controlled by specific transcription factors that lead to correct temporal and spatial expression. The activity of these factors is in turn regulated by a cascade of transcriptional regulators establishing a hierarchical regulatory scenario with a broad panel of interactions which form transcriptional regulatory networks. As a further level of regulation, the chromatin status determines the accessibility of transcription factors binding sites, thereby directly influencing transcription. Finally, post-transcriptional regulators like miRNAs contribute to a dynamic fine-tuning of mRNA abundance.

1.2 Analysis of Transcription Networks

After years of genomic research and the completion of the human genome project biology has gained a vast map containing 20,000–25,000 human genes and an even larger number of functional proteins that cooperatively regulate cellular mechanisms. However, even with the knowledge of the full genome sequence of human and other higher vertebrates the defining mechanisms are still understood to only a little extent. Epigenetics has joined genetics in the ongoing discovery comprising versatile means of regulation like the definition of chromatin status by certain histone tail modifications or the direct methylation of DNA. Both have been shown to influence gene transcription thereby providing additional layers of fine-tuning. And just recently, large parts of the genome which have previously been classified as 'junk' DNA have become the center of attention revealing additional regulatory elements like microRNAs. The analysis of these transcription networks has long been a focus of biochemical research, leading to a manifold of experimental techniques that are able to measure

regulatory dependencies even in a high-throughput manner. However, while they provide high dimensionality the pure amount of data renders a manual analysis impossible. Therefore, bioinformatic analysis methods have been developed to identify regulatory networks from experimental data, often called '*reverse-engineering*'.[36]

In general, three major questions are analyzed: Where and when is a gene (and also the appropriate protein) active, which are the regulatory factors that influence gene activation and what is the contribution of each of these individual factors to the overall regulation of genes?

The first technique used to analyze the expression of genes was northern blotting, were RNA extracts are loaded onto a gel and then detected using sequence-specific hybridization probes *e.g.* using radioactivity and photo detection. While northern blotting is in general able to detect the expression or non-expression of genes and to some extent even the amount of RNA, it has long been replaced by the more recent quantitative real-time PCR (qPCR) method. Like northern blotting qPCR detects gene expression in a sequence-specific fashion, however, using polymerase chain reaction as the functional step enables a more sensitive detection as well as more subtle graduation in how strong the gene is expressed. The main disadvantage of both techniques is that dependent on the sequence-specific probes in one reaction only one gene can be measured. This disadvantage can in some part be reduced by using the help of robots; however, the effort for measuring larger quantities is still more ore less linearly related to the number of analyzed genes. Therefore, high-throughput techniques have been developed, which are able to screen expression levels of thousands of genes at the same time using a single experiment. One of the most popular examples is microarrays, which consist of an arrayed series of thousands of microscopic spots of sequence-specific DNA oligonucleotides that are fixed to a cover slide. During a microarray experiment, labeled cDNAs derived from RNA will hybridize to complementary probes and emit light. The light intensity of each spots is dependent on the amount of DNA molecules bound, therefore providing a measure for the gene's expression. As all spots can be measured simultaneously, this leads to many thousands of data points per experiment, enabling a simultaneous analysis of all or a large fraction of the genome. More recently, next-generation sequencing is replacing microarrays as the standard technique for gene expression determination. Using next-generation sequencing the sequences of all RNA molecules in a sample can be sequenced in parallel, thereby providing a detailed snapshot of the transcriptome without any pre-design of the analyzed sequences but with the possibility to further analyze the sequence of each RNA molecule.

While the analysis of gene expression tries to answer the first question of gene activity, a further set of experimental techniques has been developed which aim to discover the regulators of gene expression, namely transcription factors and accompanying histone modifications. The first technique that analyzed the binding of proteins to DNA or RNA was the electrophoretic mobility shift assay (often called 'EMSA'). Thereby, proteins are incubated with specific short DNA sequences (*e.g.* regions from a promoter sequence) leading to a 'shift' in the resulting gel band when the protein binds to the DNA in comparison to the unbound protein. It is easily seen that this technique is not feasible to conduct genome-wide analyses, as both the protein and the DNA sequence must be determined beforehand. A major step in the analysis of regulatory networks was the invention of the chromatin immunoprecipitation (ChIP) technique. In ChIP experiments, proteins are fixed to the DNA, which is successively analyzed using either microarrays or next-generation sequencing. The advantage of ChIP in comparison to traditional methods like EMSA is the possibility to screen a full genome in one experiment without previous knowledge of the occupied DNA sites and the opportunity to perform the

experiment *in vivo*. As starting material cells from cell culture, tissue, primary cells or embryos can be used.

To answer the question of the individual contribution of found regulatory factors a number of experimental techniques have been invented. Two popular examples are the overexpression as well as the knockdown of genes. To overexpress a protein large quantities of DNA plasmids, which contain the DNA sequence of the analyzed protein fused to a potent promoter, are introduced into the cell leading to large quantities of the introduced protein. Knockdown experiments on the other hand work in the opposite direction by reducing the amount of RNA for a specific protein with the help of RNA interference (RNAi). To quantify the resulting consequences on target genes, qPCR, microarrays or next-generation sequencing are the usual methods of choice.

Hand in hand with increasing biological knowledge the mentioned experimental techniques render a simultaneous analysis of thousand of genes, their interactions and regulatory implications possible. Though, with the advent of these high-throughput techniques the need for the development, adjustment and application of reliable bioinformatic tools for their analysis has become an indispensable part of today's laboratory work. Microarrays provided one of the first areas of intense bioinformatic research, enabling not only the analysis of significant differences between different samples (*e.g.* healthy and disease tissues) but also the development of tools to cope with the problems frequently arising in high-throughput experiments. Commonly, each data measurement and mass data in particular is distorted by noise which must be dealt with to allow a reliable and unbiased analysis of the data. Therefore, normalization of experimental data as well as the correction for the high number of hypothesis test are crucial steps in every high-throughput data analysis. Further, each dataset will contain only a limited number of relevant data points, with a large fraction of uninformative data regarding the purpose of the analysis. Tools like linear modeling, clustering or the analysis of common functional annotations (*e.g.* Gene Ontology terms) are able to extract important features of the underlying datasets. Nowadays the analysis of sequencing data has become more and more important as next-generation sequencing is becoming the detection method of choice in many high-throughput experimental techniques. It is heavily based on an alignment of the resulting reads to a reference genome, which was only made possible with the increasing number of fully sequenced organisms. Further, experimental analysis will not be valid in any case, *e.g.* certain proteins cannot be analyzed with ChIP due to unspecific antibodies, and therefore bioinformatic prediction of likely regulators states a further important contribution to biochemical research. Finally, due to the high number of different experimental techniques and the complexity of the underlying biological mechanisms, an eligible analysis should therefore integrate data from not only a single but a multitude of heterogeneous sources to gather true understanding.

1.3 Transcription Networks in Cardiac Development and Disease

1.3.1 The Human Heart

The heart is the first organ to form and function during embryogenesis and starts to contract after three weeks of gestation in human. The heart is responsible for pumping blood throughout the blood vessels by repeated, rhythmic contractions. De-oxygenated blood from the body arrives through the venae cavae at the right atrium, is transported via the tricuspid valve into the right ventricle and then through the pulmonary artery into the lung, where it is re-oxygenated (Figure 3). This oxygenated blood is then

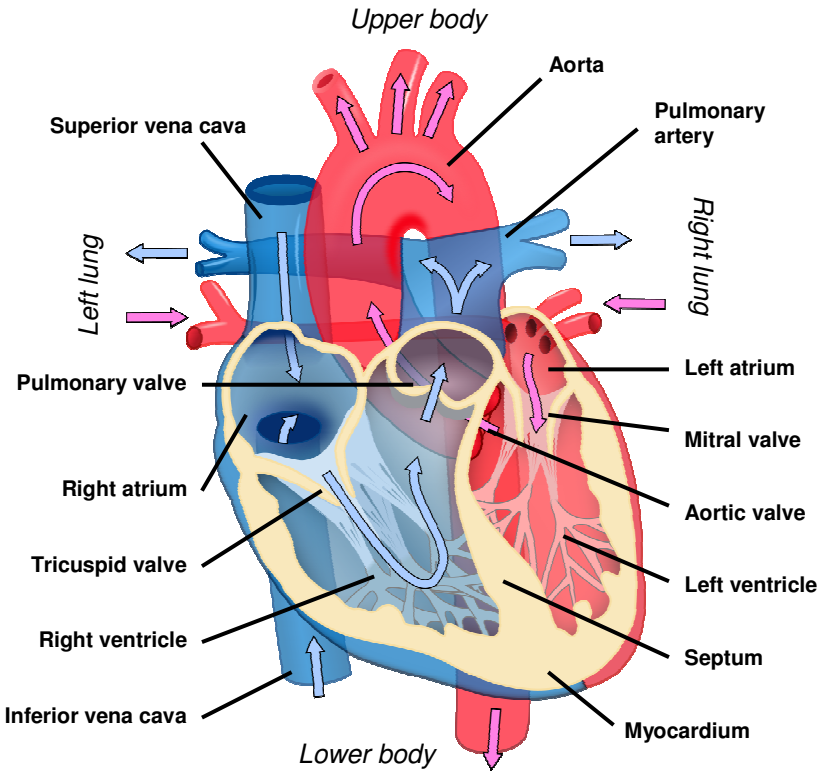

Figure 3: Schematic representation of the four-chambered human heart
Modified figure taken from Zoofari.[40]

transported to the left atrium and through the mitral valve into the left ventricle from where it is pumped through the aortic valve to the aorta and further into the body.

As a correct development is fundamental for the heart's function, it involves various molecular pathways and complex morphogenetic changes with its spatial and temporal orchestration precisely controlled by an evolutionary conserved gene program. The mammalian heart comprises a large number of different cell types, including cardiomyocytes, smooth muscle cells, endothelial cells, valvular cells and cells belonging to the conduction system.[37] These cells originate from a set of multipotent progenitor cells in the early embryonic heart field, which can be divided into two fields, the first heart field (FHF) and the secondary heart field (SHF).[38] After the heart is completely developed, the left ventricle of the four chambered heart was formed by precursor cells of the FHF, while the outflow tract, the right ventricle and most of the atria will have been formed by precursor cells of the SHF.[39]

Important stages of the development of the mammalian heart are depicted in Figure 4. After two weeks of human development, the cardiac crescent is formed, mainly by the cells of the FHF. At around day 20, a beating linear heart tube is formed with an inner layer of endocardial cells and an outer layer of myocardial cells.[41] SHF cells migrate to both ends of this heart tube and start to differentiate.[42] After four weeks of gestation, the linear tube loops into an S-shape which reflects the following heart compartments: the chamber and non-chamber myocardium, the atria, the ventricles, the outflow tract, the inflow tract and the atrioventricular canal. Growth of the cardiac chambers is achieved by increased cell proliferation during the process of 'ventricular ballooning'.[43] Around day 32, the final chambers are formed and septated which separates the blood flow into an oxygenated and a de-oxygenated system. Simultaneously, primitive valve like structures or 'cushions' are formed in

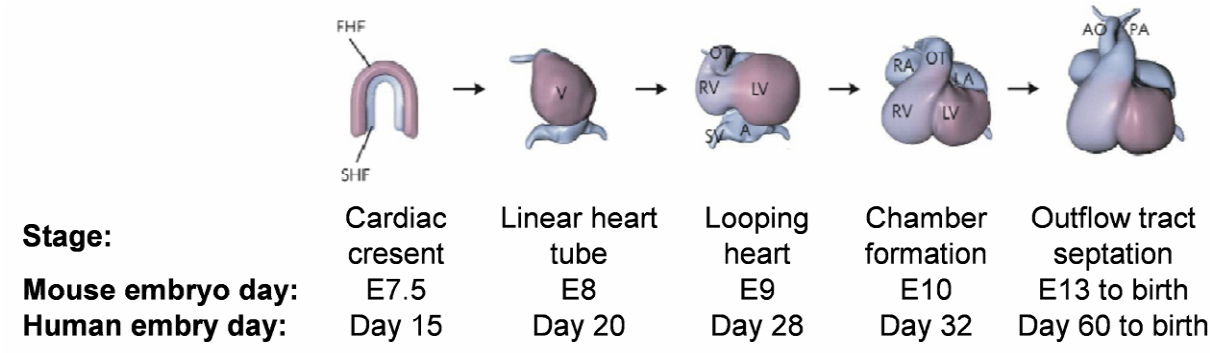

Figure 4: Mammalian heart morphogenesis
The mammalian heart is formed based on the first and secondary heart field (FHF and SHF). The left ventricle (LV) of the four chambered heart is formed by the FHF, while the outflow tract (OT), the right ventricle (RV) and the left and right atria (LA and RA) are formed by the SHF. The septation of the common outflow tract (OT) into the aorta (AO) and the pulmonary artery (PA) occurs during the outflow tract septation. Modified figure taken from Bruneau *et al.*[39]

the atrioventricular canal and the outflow tract. Finally, at around day 60, the outflow tract is septated into the aorta and the pulmonary artery. The human heart is completely developed at around 8 weeks of gestation. The cells stop to proliferate and further cardiac growth is mainly achieved by increasing the cell size (hypertrophy) rather than cell division and the cardiomyocytes withdraw from the cell cycle.[44]

1.3.2 Regulation of Heart Development

Cardiogenesis is regulated by a molecular network that comprises regulatory interactions between a multitude of transcription factors, their downstream targets and signaling pathways.[45] A set of crucial and evolutionary conserved transcription factors, comprising Gata4, Mef2 and Nkx2 factors as well as the general muscle transcription factor Srf, will be introduced in the following paragraphs.

Gata4 is one of the earliest expressed transcription factors in developing cardiac cells.[46] It belongs to the Gata family of zinc finger transcription factors, which all recognize the DNA sequence "GATA". Mutations in this gene have been shown to lead to severe heart malformations. Mice lacking Gata4 show failure in their ventral morphogenesis and heart tube formation and die before birth.[47]

'Myocyte enhancer factor 2' (Mef2) proteins are important regulators of cellular differentiation. Mef2 proteins contain both a MADS-box and a Mef2 DNA-binding domain. Four transcripts of MEF2 exist in human (MEF2A, MEF2B, MEF2C and MEF2D), which are all expressed in every stage of the developing human heart[48] and are essential for the expression of muscle-specific genes in cardiac and skeletal muscle.[49] Mice lacking Mef2a die within the first weeks of life with a range of severe heart malformations.

The homeobox protein Nkx2.5 is also expressed in the early heart and remains high in the adult heart.[50] The homologous drosophila gene called 'tinman' results in loss of heart formation during embryogenesis, suggesting a similarly important function for Nkx2.5 in human heart formation.[51] In line with this, mice lacking Nkx2.5 show severe defects in heart looping and chamber formation and die early during embryonic development.[52]

Another MADS family member is the 'serum response factor' (Srf), which is an important developmental protein, not only in the heart but in a number of tissues and participates in the

regulation of the cell cycle, apoptosis, cell growth and cell differentiation. It binds to the CArG-box motif $CC(A/T)_6GG$ in the promoters of its targets genes[53] and appears to be autoregulatory.[54] Srf is known to interact with both, positive and negative co-regulators, the most prominent being Myocardin, a smooth muscle and cardiac muscle-specific transcriptional coactivator. The interaction of Srf and Myocardin is supposed to be the inducing step for smooth muscle differentiation.[55] Mice embryos lacking Srf die during embryogenesis at the stage of gastrulation.[56]

All these factors have been shown to cooperatively regulate individual target genes. For example Nkx2.5 physically interacts with Gata4 to synergistically activate a number of downstream target genes[57] and Gata4 was shown to interact with Mef2,[58] Nkx2.5[59] and Srf.[60] In addition, these transcription factors were found to regulate each others expression, thereby potentially stabilizing their regulatory networks.[61]

However, they do not only regulate on the level of direct transcriptional control, but also indirectly by influencing the chromatin status of target genes. Mef2 proteins can act as transcriptional activators and repressors by interacting with HATs and HDACs. Further, it has been described that Srf recruits the HAT p300, possibly in conjunction with its co-factor Myocardin, which then activates target gene expression.[62] The important function of p300 in heart development is further underlined by knockout mice models which show a range of severe cardiac defects.[63] The same p300 also acetylates Gata4 itself, thereby enhancing its activating potential,[64] pointing to a high degree of interdependency. HDAC4 on the other hand was shown to repress Gata4, Mef2c, Nkx2.5 and Srf depicting the high level of interdependency between these two levels of regulation.[65]

1.3.3 Congenital Heart Disease

Congenital heart disease (CHD) is the most common birth defect with an estimated incidence of around 1% in all live births[66] and the cause for a high number of miscarriage and stillbirth.[67] The defects range from minor or even subclinical defects to complex malformations. The later have the potential to be life threatening and are therefore treated by corrective surgery which aims to restore the heart function. However, subclinical defects, even if they do not directly interfere with heart function initially, can lead to cardiovascular complications in the adult human, such as stroke or heart failure.[68]

Almost all parts of the heart can be affected, classifying the disease phenotype into three categories: septation defects, left-side obstruction defects and cyanotic heart defects.[39] Typical septation defects are the atrial septal defect (ASD), the ventricular septal defect (VSD) and the atrioventricular septal defect (AVSD). Examples for left-side obstruction defects are the aortic stenosis and an interrupted aortic arch. Cyanotic heart defects, or “blue baby syndrome”, are defects which result from the mixing of oxygenated and deoxygenated blood and cause a blue skin color. Examples are a transposition of the great arteries (TGA), and the persistent ductus arteriosus (PDA). The most common (6%) of all cyanotic defects is Tetralogy of Fallot (TOF). TOF is a complex disease and is defined of four distinct clinical features: a VSD, right ventricular hypertrophy, which is a thickening of the right ventricular walls, right ventricular outflow track obstructions, which is a narrowing at or just below the pulmonary valve, and an overriding aorta, an aortic valve of biventricular origin (Figure 5). It ultimately leads to cardiac failure with a survival rate of ~60% after four years.[69] 20-30% of all CHDs occur together with other birth defects and as parts of a syndrome, like DiGeorge syndrome or Holt-Oram syndrome.

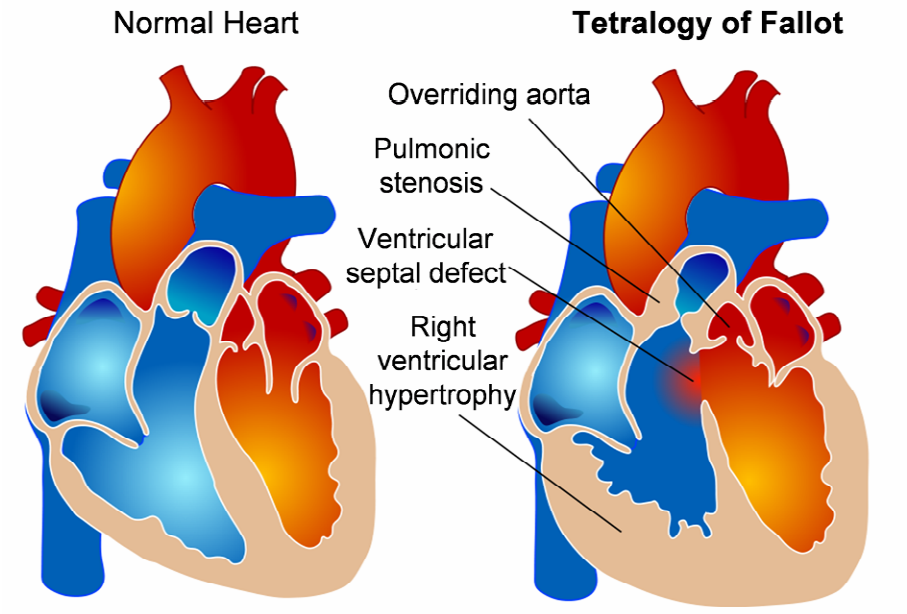

Figure 5: Tetralogy of Fallot
Schematic representation of a normal heart (left) and a heart with the 'Tetralogy of Fallot' phenotype (right) depicting the four clinical features and the mixture of oxygenated (red) and deoxygenated (blue) blood. Figure taken from Ruiz.[70]

Using linkage analysis in nonsyndromic families several genes mutated in human CHDs have been identified. Among these are ACTC1 (ASD),[71] GATA4 (ASD),[72] JAG1 (TOF),[73] MYH6 (ASD),[74] MYH11 (PDA),[75] and ZIC3 (TGA).[76] In line with the regulation of heart development by genetic pathways, some mutations in different genes were found to result in the same or very similar disease phenotype, based on the disturbance of the same specific regulatory pathway. Deletion of either BMP, ALK2, ALK3, ALK6 or SMAD were *e.g.* shown to result in the same cardiac defect.[77] Further, a genome-wide expression analysis in human conducted by Kaynak *et al.*[78] revealed genetic changes that are correlated with distinct congenital malformations.

However, in general only a minor fraction of CHDs are the result of monogenic disorders with a clear Mendelian inheritance. Instead, most CHDs display variable expressivity and penetrance pointing to a multifactorial and multigenetic basis which is in most cases poorly understood. In human and mice similar mutations can cause a variety of phenotypes from one family, individual or inbred strain, respectively, to another. Heterozygous mutations of NKX2.5 in human can *e.g.* lead to such diverse abnormalities as ASD, VSD, Ebstein's anomaly of the tricuspid valve, AV block or TOF, either alone or in combinations.[79] A similar situation exists for the T-box factor TBX5, in which heterozygous mutations cause a variety of CHDs in the context of Holt-Oram syndrome.[80] In other cases the same mutation in two individuals might lead to a disease in one individual, while the other seems to be unaffected. Possible explanations for this reduced penetrance are buffering by a second allele or other TFs of the cardiac pathway.[81,82] Additionally, the disease manifestation may vary due to stochastic events of unknown nature or further parameters comprising genetic modifiers and environmental influences. For instance, maternal diabetes and obesity have been shown to promote CHDs as well as alcohol, anti-depressants, herbicides or infections during early pregnancy.[83-85] A valid suggestion is therefore, that CHD patients assemble multiple genetic and non-genetic factors which reduce the properties of the networks to buffer individual disturbances and finally lead to cardiac malformation.[86]

Thus, disturbances in genetic pathways might be based on environmental influences, genetic mutations or a combination of both.[77]

In addition to the multifactorial and multigenetic basis of CHD the symptom severity of cardiac defects also depends on the type of mutation. Some missense mutations result in non-functional proteins, whereas others may lead to altered properties of unknown nature.[87] As an example, certain mutations in Tbx5 abolish its binding to DNA[88] while others influence collaborations with other proteins.[89] Secondary adaptation processes are a further factor that obstructs the identification of genes causal for CHDs. They have no direct genetic origin but are caused by the need to maintaining the heart function during its development.

This suggests that the regulatory context of TFs plays an important role and their function must be viewed in the framework of transcriptional networks, including the interplay between different TFs as well as epigenetic factors such as histone modifications or other regulatory factors such as microRNAs. However, while single regulatory dependencies have been deeply evaluated the amount of current knowledge on overall regulatory networks that drive correct and incorrect heart development is still very premature. In the same line, the possibilities to diagnose certain CHD *in utero* as well as advances in surgical techniques have considerably improved life expectances and quality for children born with CHD. However, preventing CHD and elucidating their causative factors are still major goals.[86] Thus, the analysis of transcriptional networks and their disturbances is required to discover novel targets for diagnostics and therapeutics.

1.4 Publications

Several parts of this study have been published. Much of the work described in section 3.1 has been published in the paper "The Cardiac Transcription Network Modulated by Gata4, Mef2a, Nkx2.5 and Srf, Histone Modifications and MicroRNAs", *PLoS Genetics* (in press). The work presented in section 3.3 has been published in the paper "Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes", *Molecular Biosystems* (2008). Further, a manuscript containing the work described in section 3.2 with the title "Dynamics of histone modifications and transcription factor binding during cardiac maturation in mice" is currently in preparation. The CArdiovascular Regulatory INteraction database described in section 3.4 was distributed as the dissemination database of the HeartRepair EU project.

2. Material and Methods

Bioinformatic analysis has become a major component in every high-throughput analysis ranging from the *in vivo* monitoring of TFBS to expression changes resulting from perturbations of regulatory networks. In the following sections, experimental techniques used to gather these high-throughput datasets are highlighted (section 2.1) and the different datasets that were investigated throughout this study are described (section 2.2). Finally, the bioinformatic approaches that were used to analyze these datasets are depicted in detail (section 2.3).

2.1 Experimental Methods

This study integrates a number of advanced experimental techniques which have been applied to study the cardiac regulatory network. These comprise the qPCR technique which measure amounts of DNA or reverse-transcribed mRNA in a sample (section 2.1.1), the monitoring of protein-DNA interaction using ChIP (section 2.1.2) either followed by array detection (section 2.1.3) or the more recent next-generation sequencing (section 2.1.4). In addition, RNAi as a method to specifically reduce the amount of specific mRNAs in a cell is described (section 2.1.5).

2.1.1 Quantitative Real-Time-PCR (qPCR)

Quantitative real-time PCR follows the general principle of polymerase chain reaction which amplifies DNA molecules but includes an additional step of target quantification after every cycle of amplification. This allows an estimation of the amount of DNA which was initially present in the sample. The quantification is commonly done using fluorescent reporter probes (primers) that detect only the DNA matching the probe sequence. The emitted light intensity is linearly correlated with the amount of amplified DNA in each cycle. To get a reliable estimate for the initial amount of DNA, the time point were the fluorescence significantly exceeds the background for the first time (called ‘ct’ or ‘cp’) is used.

However, it is uncommon to quantify the absolute amount of sample DNA, as the measured ct/cp value is influenced by a number of factors, *e.g.* the efficiency of the reverse transcriptase. Instead, a relative value between the measured DNA sample and an internal control, often a ‘housekeeping’ gene with expected stable expression or a reference DNA sample, is used for normalization. As the PCR reaction doubles the amount of target DNA in every cycle the formula

$$relative\ amount = 2^{-\Delta ct} = 2^{-\left(ct_{control} - ct_{sample}\right)} \quad ,$$

where $ct_{control}$ and ct_{sample} refer to the ct value of the control and the target DNA, respectively, is used to calculate the relative amount of DNA. This formula is only valid if the efficiency of the PCR is close to 2, which must be validated for every experiment. More accurate methods use a statistically more stable control, like the geometric mean between a set of housekeeping genes as suggested by Vandesompele *et al.*[90] which was applied in the qPCR analysis of patient data. In addition, every qPCR experiment should consist of several replicates to assess statistical variability.

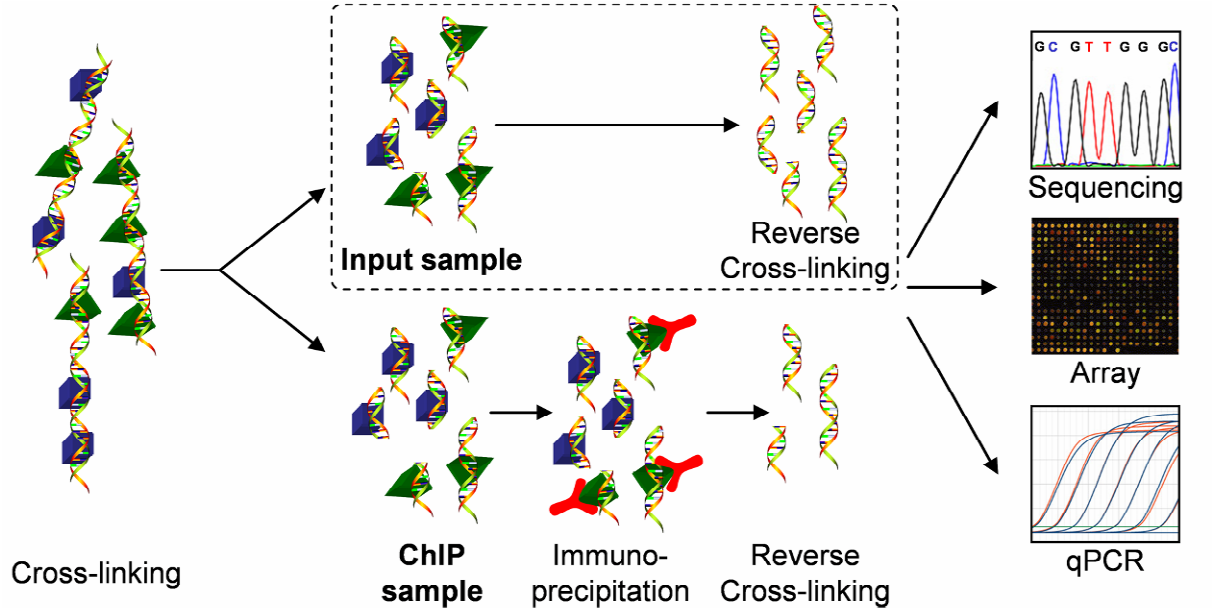

Figure 6: Schematic representation of a chromatin immunoprecipitation (ChIP) experiment
The part included in the dashed box might be left out in certain ChIP experiments.

2.1.2 Chromatin Immunoprecipitation (ChIP)

Chromatin immunoprecipitation is a high-throughput experimental technique to screen for genomic regions covered by proteins. It consists of several steps which are depicted in Figure 6. The primary step in every ChIP experiment is called cross-linking and involves the chemical anchoring of the proteins to the DNA, usually achieved via formaldehyde. In this step, whole protein complexes are also cross-linked, allowing a measurement of genomic sites for proteins that do not directly bind to DNA. After cross-linking, the chromosomal DNA is sheared to fragments of desired size, usually using sonication (application of ultrasound). The next step is the enrichment of those DNA samples that are bound to the protein of interest. This is performed using antibodies specifically directed against the protein in a procedure called immunoprecipitation. All DNA fragments not bound to the protein will be washed away, leaving the enriched 'ChIP sample'. In many experiments, an additional sample is prepared which is not immunoprecipitated to measure the experimental background. This sample is typically called 'Input sample' or 'Input' and is required for many bioinformatic ChIP analysis algorithms. The last step in the ChIP protocol is a reset of the protein-DNA cross-linking, which is called reverse cross-linking, and the purification of the DNA to remove the proteins. If the resulting ChIP material is too low, an additional amplification step is required; however, if possible this should be avoided due to potential amplification biases.

Afterwards, the DNA from the resulting ChIP sample must be determined to predict DNA regions of protein binding. Depending on the methods to detect the DNA, the ChIP experiment is called 'ChIP-chip'/'ChIP-on-chip' (ChIP followed by microarray analysis) or 'ChIP-seq' (ChIP followed by next-generation sequencing). In addition, it is possible to use quantitative real-time PCR to determine if specifically selected DNA regions are enriched in the ChIP sample in comparison to the Input sample. This is sometimes called 'ChIP-qPCR'.

2.1.3 ChIP Followed by Microarray Analysis (ChIP-chip)

Depending on the used array several slightly different techniques to measure ChIP enrichment data using array analysis have been proposed. However, the most common form uses fluorescence labeling (usually the cyanine dyes Cy3 for green and Cy5 for red) of the ChIPed DNA, usually Cy3 for the Input and Cy5 for the ChIP sample, and hybridization to a single or multiple microarrays. The fluorescence signal is subsequently measured using laser detection and relative intensities between the two dyes are then used to determine DNA fragments enriched in the CHIP experiment. Ideally, one or multiple arrays represent the entire genome of the organism in form of overlapping fragments (tiling array). However, due to the large genome sizes especially of higher vertebrates, those full-genome tiling arrays are only available at high monetary costs. Alternatively, custom designed arrays are available that span certain regions of interest, *e.g.* promoters, for a selected number of genes. The spacing of the probes on the array as well as the size of the DNA fragments after shearing determines the resolution of the identified protein-DNA binding sites.

Using ChIP-chip experiments large quantities of genomic regions can be analyzed for protein abundance, which enables a detailed study of transcriptional regulatory networks. The disadvantage of ChIP-chip experiments, however, is the need to design a fixed tiling and to reduce the regions under study to those with potential regulatory input which can leave important regions undetected. Further, only DNA fragments that are unique in the genome are spotted on the array which misses highly repetitive regions, genomic elements which have already been found to contain a number of regulatory sites.[91,92] In addition, the microarray hybridization is not very sensitive and requires high amounts of ChIP material which often entails additional amplification.

2.1.4 ChIP Followed by Next-Generation Sequencing (ChIP-seq)

Nowadays, next-generation sequencing has replaced microarray hybridization as the common method to detect the ChIPed material. A number of next-generation sequencing techniques exist with the most popular being the '*sequencing-by-synthesis*' and the '*pyrosequencing*' approaches. S*equencing-by-synthesis*, a techniques that is implemented in the 'Illumina Genome Analyzers' and which is based on the original Sanger sequencing, incorporates a preliminary cluster generation step, where the genomic DNA samples are fused to specific adapter sequences, which are then placed randomly on the surface of a flow cell. An amplification step increases the number of identical sequences locally generating several million of dense clusters of DNA. The sequencing is subsequently performed in cycles using labeled dNTPs which hybridize to the DNA one nucleotide per cycle. The last added nucleotide of each fragment is identified using laser excitation and image capturing of the emitted fluorescence. The process stops after 36 cycles resulting in reads with an exact length of 36 bp, yet, machines with 72 and more cycles are available. In contrast, *pyrosequencing*, which is implemented in the 'Genome Sequencer FLX' by Roche 454 Life Science, is based on detecting the activity of a DNA polymerase with another chemiluminescent enzyme. The polymerase adds complementary nucleotides to the DNA fragment and emits light dependent on the nucleotide. The sequence of emitted lights is captured by a camera and retranslated into reads with a current average length of 250 bp.

According to recent studies, next-generation sequencing of ChIP material appears to be much more sensitive than ChIP-chip.[93,94] Further, it removes the need to reduce the examined genomic regions and therefore allows a more unbiased analysis of transcriptional networks by measuring also regions that are thought unlikely to comprise regulatory protein binding. Though, like ChIP-chip experiments,

ChIP-seq might fail to accurately measure protein abundance in repetitive regions as these will in most cases result in reads that can be mapped to multiple genomic positions. Extending the lengths of reads using more recent sequencing approaches might reduce this problem.

2.1.5 RNA Interference (RNAi)

ChIP does reveal likely regulators for gene expression if these bind in the proximity to known transcriptional start sites. However, it does not reveal whether the TF under study is an activator, a repressor or non-functional for a certain target gene. Consequently, to access the genome-wide regulatory implications of a TF, it is necessary to couple a ChIP experiment with a transcriptome analysis. While measurement of a panel of possible target genes under wildtype conditions can give information on the overall effect of the TF on transcription, it doesn't reveal the effect of the TF for each gene individually. To circumvent this lack of information cells can be studied under conditions were the TF is inactive or absent. A common experimental method is the knockdown of the TF using RNA interference (RNAi). RNAi is a cellular mechanism that is intermingled with the miRNA pathway and plays a role in post-transcriptional regulation and defense against viral infections. It uses the aforementioned RISC complex (section 1.1) to specifically degrade target mRNA molecules that fit to a regulatory double-stranded short interference RNA (siRNA). In siRNA experiments these siRNA molecules are artificially introduced into the cell via transfection and lead to the desired knockdown of the target transcript. The expression profile of the cell can subsequently be accessed using appropriate methods like qPCR for single genes or microarray or next-generation sequencing for genome-wide expression analysis. Crucial for every siRNA experiment is the additional measurement of a non-specific siRNA (often called 'siNon') to assess consequences on the cellular transcription profile that are caused by the RNAi experiment itself and not the induced siRNA.

2.2 Analyzed Datasets

All experiments analyzed in this study were conducted to study individual components of the transcriptional regulatory network of the vertebrate heart. The first analysed datasets monitor the binding of transcription factors and histone modifications in a steady-state cell culture model using ChIP-chip (section 2.2.1) combined with microarray gene expression data of wildtype and corresponding siRNA knockdown cells (section 2.2.2). To analyze found regulatory implications in more detail, a subset of these experiments was repeated using genome-wide ChIP-seq data of the TF Srf an the histone modification H3ac (section 2.2.3). Results were validated and further enhanced using ChIP-qPCR experiments measuring the enrichment of a number of regulatory factors in a mouse time-series (section 2.2.4). Finally, transcription networks obtained in cell culture and mouse models were studied further using qPCR gene expression data from patient with congenital heart disease (section 2.2.5).

All experiments were performed by members of the laboratory of Prof. Dr. Silke Sperling, Max Planck Institute for Molecular Genetics, Department Lehrach.

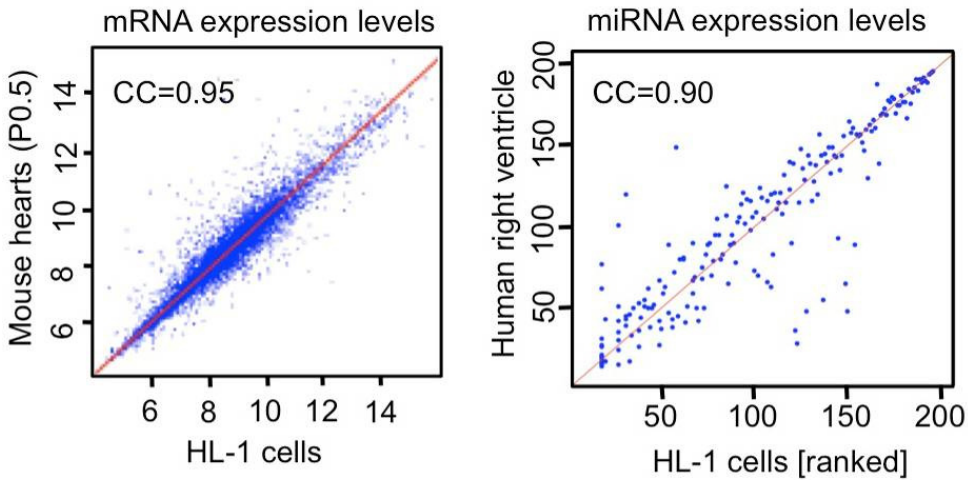

Figure 7: Comparing HL-1 cells to human and mouse hearts
(left) Gene expression levels obtained from HL-1 cells and P0.5 of C57/BL6 mouse heart. (right) Rank-transformed miRNA expression levels in HL-1 cells and human right ventricle. The Pearson correlation coefficients are indicated in the upper left corner.

2.2.1 ChIP-chip Data for Cardiac TFs and Histone Modifications in Cell Culture

As an initial step in the study of cardiac regulatory networks the binding of key regulatory transcription factors to promoters of target genes was monitored using chromatin immunoprecipitation followed by array detection. Experiments were performed using HL-1 cells for the four DNA-binding TFs Gata4, Mef2a, Nkx2.5 and Srf, which play pivotal roles for the differentiation, maturation and homeostasis of cardiomyocytes.

The cardiomyocyte cell line HL-1 was used because it is a feasible model to study cardiomyocytes, as mRNA as well as miRNA expression profiles are highly comparable to postnatal mouse hearts (Pearson correlation coefficient of 0.95, Figure 7 left) and human right ventricle (Pearson correlation coefficient of 0.90, Figure 7 right). Each ChIP experiment was performed using four replicates, two biological (two independent pools) and two technical. The ChIP materials as well as an Input control were subsequently labeled with a fluorescent dye (Cy5 for the ChIP samples and Cy3 for the input control) and hybridized to two specifically designed 385 NimbleGen arrays according to NimbleGen standard procedure. The design of these two arrays, which was performed by Tammo Krüger, a former member of our group, was based on a broad panel of muscle relevant data sources to sufficiently represent regulatory sites in enhancer and promoter regions of all known expressed skeletal, smooth and cardiac muscle relevant genes in human and mouse (Table 1). The arrays represent 89 Mb of the mouse genome mm8 (NCBI m37) associated to 12,625 transcriptional start sites and contained 740,000 probes with a tiling of 110 bp (50-60 bp gab between the probes). They included conserved regions (based on PhastCons[95] score threshold of 0.2 – section 2.3.8) within 10 kb upstream, the full sequence 2 kb upstream and the first exon and intron of the corresponding transcript. The resulting array data consisted of two different measurements per experiment, one for the Input and one for the ChIPed material.

In addition, ChIP-chip data regarding the four histone modifications H3K9K14ac (H3ac), H4K5K8K12K16ac (H4ac), H3K4me2 and H3K4me3 was used. These four histone modifications were described to promote an open chromatin state,[5,96-99] and were previously analyzed in our own group also using ChIP-chip techniques and linear modeling.[16]

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.*[100]	2,546
Symatlas human atrioventricular node – A/B[101]	2,399 / 2,399
Symatlas human cardiac myocytes – A/B[101]	4,786 / 3,981
Symatlas human heart – A/B[101]	3,391 / 3,978
Symatlas human skeletal muscle – A/B[101]	1,889 / 1,761
Symatlas human smooth muscle – A/B[101]	5,296 / 5,237
Symatlas mouse heart	1,665
Symatlas mouse skeletal muscle	1,793
Transcripts expressed in mouse hearts – Tabibiazar *et al.*[102]	132
All transcription factors listed in TRANSFAC[103] as of Jan 2005	2,236

Table 1: Sources considered for array design

2.2.2 Microarray Expression Data for Wildtype and RNAi Knockdown

To study if the binding of transcription factors observed in the ChIP experiments would influence the expression of associated genes, genome-wide expression array analysis of contracting HL-1 cardiomyocytes was carried out using Illumina microarrays. All measured transcripts were represented by several probes on an array to gain higher statistical power for detecting differential gene expression. As the number of probes and measurements exceeded the size of a single array, multiple arrays were used.

First, steady-state expression data of untreated HL-1 cells was measured in two replicates. Then, HL-1 cells were treated with siRNAs against a single TF, leading to a major reduction of the quantity of each TF in each cell. Two different siRNAs per TF were used, each in duplicates, leading to 4 replicate experiments per TF. Finally, the cells were treated with an unspecific siRNA (siNon) to measure any bias introduced by the siRNA treatment itself, again using duplicates.

2.2.3 ChIP-seq Data for Srf and Histone 3 Acetylation in Cell Culture

In an attempt to confirm and further investigate results obtained from the analysis of the ChIP-chip data, additional ChIP experiments were conducted now followed by next-generation sequencing. Two independent ChIP samples were profiled. After chromatin immunoprecipitation was performed, DNA fragments bound by Srf or modified with H3ac in HL-1 cells were sequenced using the next-generation single-end sequencing technology of the Illumina Genome Analyzer which sequences reads of 36 bp in length using the *sequencing-by-synthesis* approach. Analysis of the resulting images and successive base calling was done using the open source Firecrest and Bustard applications[104] (Solexa pipeline 1.4.0). The sequencing of the small RNA libraries resulted in 6,967,318 and 8,364,328 sequence reads obtained in the Srf and H3ac ChIP-seq experiment, respectively.

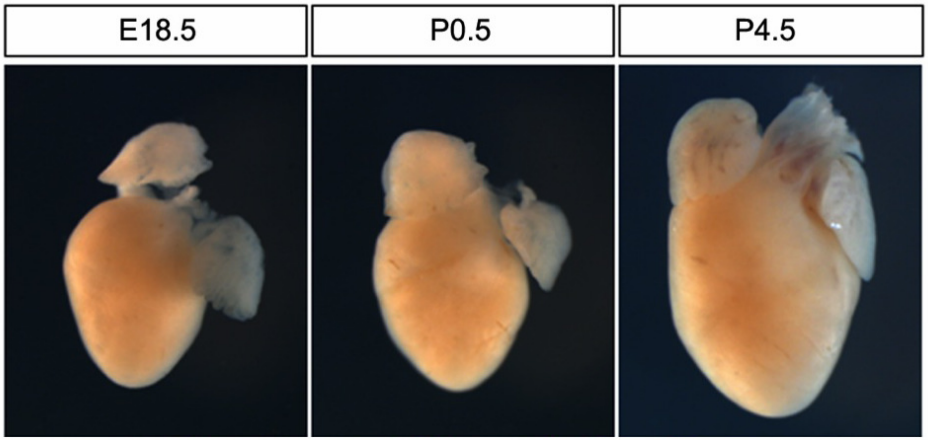

Figure 8: Mouse hearts of three developmental stages
Three stages of cardiomyocyte maturation around birth. From fetal to the postnatal stage the heart heavily increases in size. The pictures are shown by courtesy of Jenny Schlesinger.

2.2.4 Time Series ChIP-qPCR Data of TFs and Histone Modifications in Mouse Hearts

After the analysis of TF binding and histone modifications was performed in a steady-state cell culture model, further experiments were conduct to investigate the impact of histone modification on Srf regulated gene expression in more detail in a time series manner. Experiments were performed in mouse hearts of three developmental stages, one prenatal stage (E18.5, meaning 18 ½ days after fertilization) and two postnatal stages (P0.5 and P4.5, meaning ½ and 4 ½ days after birth). From the fetal to the postnatal stage, the heart adapts to the body circulation and cardiomyocytes mature. During this process the heart increases in size (Figure 8), the cells elongate, myofibrils align and cell-cell contacts become bipolar.[105] Srf, the two histone modifications H3ac and H3K4me2 as well as the histone acetyltransferase p300, which was proposed to couple Srf-binding to H3ac enrichment, were measured using chromatin immunoprecipitation followed by quantitative real-time PCR on a set of selected promoter regions. QPCR was used because it allows a sensitive detection of ChIP enrichment changes even for the very small tissue amounts that can be gathered from mouse hearts in these early stages. Though, using qPCR to detect the amount of ChIPed DNA requires the definition of a predefined set of regulatory regions that should be analyzed, as the qPCR reaction depends on sequence specific primers. To ensure high comparability to the results gathered in cell culture the selection of regions with potential regulatory influence was based on results from the ChIP-chip/seq analysis gathered in HL-1 cardiomyocyte cells. The selection process was performed in three steps:

First, genomic regions containing overlapping ChIP peaks of at least two of H3ac, H3K4me2 and Srf were determined (a description of the ChIP peak calling procedure and its application to the ChIP data refer is given in sections 2.3.5, 2.3.6 and 3.1.2). Requiring a maximal distance of 500 bp between the peaks' mid points resulted in 2,484 selected regions. The genomic position of each selected region was subsequently defined based on the start of the first enclosed peak to the end of the last enclosed peak thereby spanning all factors of interest fully (Figure 9).

In a second step, all selected regions were associated to the most proximal gene if they lay inside the transcribed regions or not more than 10 kb upstream using mouse gene annotations from Ensembl[106] (version 55). To reduce the number of selected regions to those interesting for cardiac development

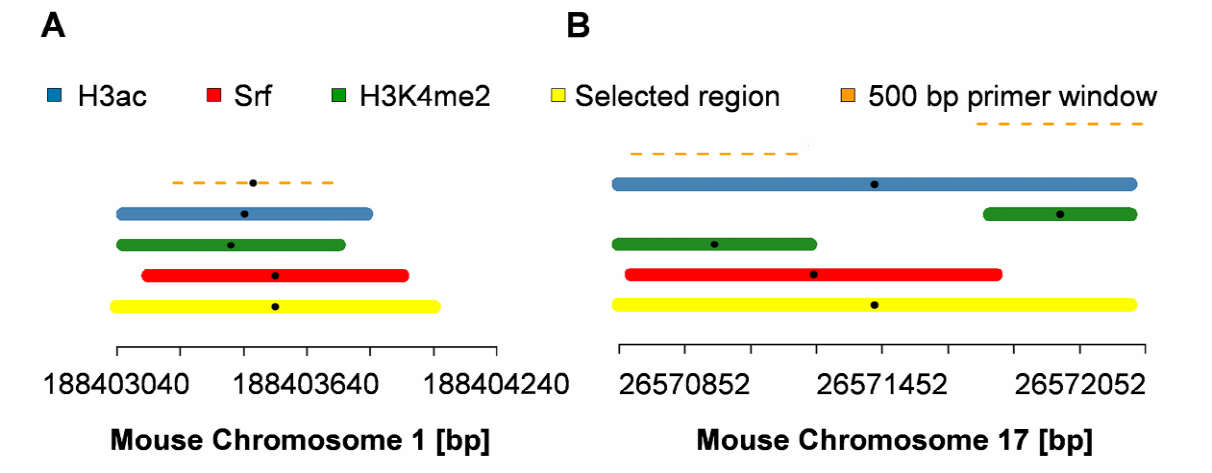

Figure 9: Region selection for the ChIP-qPCR measurement
Two examples of selected regions based on ChIP peaks gathered in HL-1 cells. The selected regions (yellow) span all overlapping individual ChIP peaks (blue, green and red for H3ac, H3K4me2 and Srf, respectively). Fixed windows of 500 bp length (orange) were positioned in the middle of the ChIP peaks for primer design. **(A)** A single 500 bp primer window was associated with this selected region on chromosome 1. **(B)** Two primer windows were used to span all interesting ChIP peaks associated to this selected region on chromosome 17.

regions were filtered that could be associated to genes with heart or muscle annotation background using a combination of resources including experiments, literature search, OMIM and GO annotation.

In a third step, additional regions were manually included that were bound by Srf in ChIP-chip/seq and were associated to genes which are important for heart and muscle development but failed to be automatically annotated (possibly due to insufficient GO annotation) or were found to be interesting based on our own publication.[107]

After the selection of regulatory regions for the ChIP-qpCR experiment, primer pairs were designed that uniquely represent the regions of interest. Therefore, a fixed size window of 500 bp length was positioned inside each selected region using the middle of all ChIP H3ac, H3K4me2 and Srf peaks that were enclosed in the selected region. If a single 500 bp window did not cover all individual ChIP peaks, additional non-overlapping 500 bp regions were introduced manually (Figure 9 B). Using these 500 bp windows, sequence specific primer pairs were designed and regions without unique primer pairs were discarded from the subsequent analysis. This selection process resulted in 191 regions with regulatory context in the proximity of heart developmental genes that were bound by histone modifications and/or Srf in HL-1 cells.

Finally, ChIP-qPCR was performed for every selected region with samples of mouse hearts of each individual time point (E18.5, P0.5 and P4.5) and each measured TF and histone modification (Srf, p300, H3ac and H3K4me2,). To adjust for different initial DNA concentrations, an Input control was measured. After the qPCR measurement, replicates with very high standard deviations were inspected manually to remove potential outliers and measurements that completely failed in all three replicates were excluded from the subsequent analyses. At last, the relative amount of ChIP enrichment in each region for each time point and measured TF and histone modification was calculated using the measured Input.

2.2.5 qPCR Expression and Phenotype Data of Patients with Congenital Heart Disease

Finally, the relevance of the data obtained using cell culture and mouse hearts were analyzed in human hearts. Gene expression was measured for a selected set of cardiac relevant genes using qPCR. As a further contribution to the range of this study the additional analysis of breakdowns of cardiac transcription networks was studied by integrating a large number of human patients with a panel of congenital heart disease supplemented by a control group of healthy individuals. In total 190 human cardiac ventricular and atrial tissue samples from patients with different cardiac malformations were collected, enabling the selection of a balanced patient population and allowing the separation of disease- or tissue-specific expression patterns. All cardiac samples were obtained from the 'German Heart Center' during cardiac surgery with ethical approval by the institutional review committee and informed consensus of the patients or their parents. As the control group healthy human heart samples from non-transplanted hearts were taken. To retrieve phenotypic information a clinical characterization comprising 250 features of morphological, hemodynamic and therapeutic information was collected for every analyzed patient using the *d*-matrix database[108] for detailed analysis and visualization.

To characterize the transcript patterns of the patient samples, a set of 42 genes was selected based on a previous genome-wide study by Kaynak *et al.*[100] as well as literature research. For these genes qPCR primers were designed to determine expression levels. Table 2 shows the list of all measured genes and their Ensembl Gene IDs in human and mouse. In addition to the 42 genes, four genes with expected stable expression in the diseased individuals were measured additionally and the geometric mean of the three most consistent genes was calculated for each sample according to the method suggested by Vandesompele *et al.*[90] This mean was then used to calculate relative mRNA amounts and the four genes were excluded from subsequent analyses.

2.3 Bioinformatic Methods

In the previous section the range of different experiments that have been integrated into this study were described. In line with this, a number of bioinformatic tools had to be implemented to analyze their results and derive meaningful hypothesis for future studies of cardiac transcription networks. In the following, these tools are described in detail. A technique frequently applied was linear modeling (section 2.3.1), which is a versatile and very flexible tool for the statistical analysis of data. Just as important, the normalization methods used to remove biases from the high-throughput datasets (section 2.3.2) as well the implemented measures to assess statistical dependency and derive clusters of co-expressed genes (section 2.3.3) are described. As most of the analyses include multiple testing procedures its correction is explained (section 2.3.4) and the pipelines to derive enriched binding sites from ChIP data are elucidated for ChIP-chip (section 2.3.5) and ChIP-seq (section 2.3.6). Finally, additional tools that have been applied in this study, like Gene Ontology (GO) term enrichment analysis (section 2.3.7), the prediction of transcription factor binding sites either using known PWMs (section 2.3.8) or *de novo* (section 2.3.9) and relational databases (section 2.3.10) are illustrated. If not mentioned otherwise, all bioinformatic analyses have been performed using scripts implemented in Perl, R[109] and the Bioconductor[110] packages.

Gene	Human Ensembl ID	Mouse Homolog	Mouse Ensembl ID	Description
ACTA1	ENSG00000143632	Acta1	ENSMUSG00000031972	Actin, alpha skeletal muscle
ATP2A2	ENSG00000174437	Atp2a2	ENSMUSG00000029467	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2
BMP2	ENSG00000125845	Bmp2	ENSMUSG00000027358	Bone morphogenetic protein 2 precursor
BMP4	ENSG00000125378	Bmp4	ENSMUSG00000021835	Bone morphogenetic protein 4 precursor
CITED2	ENSG00000164442	Cited2	ENSMUSG00000039910	Cbp/p300-interacting transactivator 2
CPT1B	ENSG00000205560	Cpt1b	ENSMUSG00000078937	Carnitine O-palmitoyltransferase I, muscle isoform
DPF3	ENSG00000205683	Dpf3	ENSMUSG00000021221	Zinc-finger protein DPF3
GATA4	ENSG00000136574	Gata4	ENSMUSG00000021944	GATA-binding factor 4
GATA6	ENSG00000141448	Gata6	ENSMUSG00000005836	GATA-binding factor 6
HAND1	ENSG00000113196	Hand1	ENSMUSG00000037335	Heart- and neural crest derivatives-expressed protein 1
HAND2	ENSG00000164107	Hand2	ENSMUSG00000038193	Heart- and neural crest derivatives-expressed protein 2
HEY1	ENSG00000164683	Hey1	ENSMUSG00000040289	Hairy/enhancer-of-split related with YRPW motif 1
HEY2	ENSG00000135547	Hey2	ENSMUSG00000019789	Hairy/enhancer-of-split related with YRPW motif 2
HIF1A	ENSG00000100644	Hif1a	ENSMUSG00000021109	Hypoxia-inducible factor 1 alpha
HOP	ENSG00000171476	Hopx	ENSMUSG00000059325	Homeodomain-only protein
IRX4	ENSG00000113430	Irx4	ENSMUSG00000021604	Iroquois-class homeodomain protein
MEF2A	ENSG00000068305	Mef2a	ENSMUSG00000030557	Myocyte-specific enhancer factor 2A
MEF2C	ENSG00000081189	Mef2c	ENSMUSG00000005583	Myocyte-specific enhancer factor 2C
MYH6	ENSG00000197616	Myh6	ENSMUSG00000040752	Myosin-6 (Myosin heavy chain 6)
MYH7	ENSG00000092054	Myh7	ENSMUSG00000053093	Myosin-7 (Myosin heavy chain 7)
MYL2	ENSG00000111245	Myl2	ENSMUSG00000013936	Myosin regulatory light chain 2
MYL7	ENSG00000106631	Myl7	ENSMUSG00000020469	Myosin regulatory light chain 7
MYOCD	ENSG00000141052	Myocd	ENSMUSG00000020542	Myocardin
NKX2.5	ENSG00000183072	Nkx2.5	ENSMUSG00000015579	Homeobox protein Nkx-2.5
NPPA	ENSG00000175206	Nppa	ENSMUSG00000041616	Atrial natriuretic factor precursor
NR2F1	ENSG00000175745	Nr2f1	ENSMUSG00000069171	COUP transcription factor 1
NR2F2	ENSG00000185551	Nr2f2	ENSMUSG00000030551	COUP transcription factor 2
PIPPIN	ENSG00000172346	Csdc2	ENSMUSG00000042109	Cold shock domain-containing protein C2
PLOD1	ENSG00000083444	Plod1	ENSMUSG00000019055	Procollagen-lysine,2-oxoglutarate 5-dioxygenase 1 precursor
RARA	ENSG00000131759	Rara	ENSMUSG00000037992	Retinoic acid receptor alpha
RXRA	ENSG00000186350	Rxra	ENSMUSG00000015846	Retinoid X receptor alpha
SMAD4	ENSG00000141646	Smad4	ENSMUSG00000024515	Mothers against decapentaplegic homolog 4
SMAD6	ENSG00000137834	Smad6	ENSMUSG00000036867	Mothers against decapentaplegic homolog 6
SRF	ENSG00000112658	Srf	ENSMUSG00000015605	Serum response factor
TAGLN	ENSG00000149591	Tagln	ENSMUSG00000032085	Transgelin
TBX20	ENSG00000164532	Tbx20	ENSMUSG00000031965	T-box transcription factor TBX20
TBX5	ENSG00000089225	Tbx5	ENSMUSG00000018263	T-box transcription factor TBX5
TGFB2	ENSG00000092969	Tgfb2	ENSMUSG00000039239	Transforming growth factor beta-2 precursor
TNNC1	ENSG00000114854	Tnnc1	ENSMUSG00000021909	Troponin C
TNNI3	ENSG00000129991	Tnni3	ENSMUSG00000035458	Troponin I
VEGF	ENSG00000112715	Vegfa	ENSMUSG00000023951	Vascular endothelial growth factor A precursor
ZFPM2	ENSG00000169946	Zfpm2	ENSMUSG00000022306	Zinc finger protein multitype 2

Table 2: Genes screened in the patient analysis
Genes selected for the cardiac gene set. Ensembl IDs are based on Ensembl version 48. Assignment of homologous mouse genes was taken from Ensembl.

2.3.1 Regression Analysis and Linear Modeling

Linear regression models and analysis of variance (ANOVA) models are powerful statistical tools and can be applied to a number of real-world problems including bioinformatics. The aim of any regression analysis is the explanation or the modeling of the relationship between a response or observed variable Y (*e.g.* the expression of a gene) and one or more predictive variables X_1, ..., X_k (*e.g.* the expression of likely regulators). In this study, linear modeling approaches have been applied to study a number of problems, including the prediction of regulators as in the analysis of the regulatory impact of p300 and Srf on changes in H3ac enrichment using time-series data (section 3.2.3 to 3.2.5) or the differential expression of genes in patients belonging to phenotypic subgroups (section 3.3.4).

In regression analysis, the data used for the regression is usually present in form of a vector for the response variable and a data matrix for the predictive variables like

$$\left(\begin{array}{c|cccc} y_1 & x_{11} & x_{12} & \vdots & x_{1k} \\ y_2 & x_{21} & x_{22} & \vdots & x_{2k} \\ \dots & \dots & \dots & \ddots & \\ y_n & x_{n1} & x_{n2} & & x_{nk} \end{array}\right) \quad ,$$

where n is the number of observations. The most general way to model this dependency would be

$$Y = f\left(X_1, X_2, ..., X_k\right) + \varepsilon \quad ,$$

where f is an unknown function and ε represents the individual error or noise that is found in every measurement. However, as the real function f could be infinite dimensional, it is unfeasible to estimate such a function. Therefore the form of f must be restricted to be able to estimate a model. The linear model, one of the most prevalent methods of regression analysis, assumes a linear form of f.

The Linear Model, Linear Regression Models and ANOVA

Linear models are models of the restricted linear form

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k + \varepsilon \quad ,$$

where β_0 to β_k are the unknown parameters of the model which are to be estimated. β_0 is called the intercept, which is introduced to represent a general baseline not accounted by the predictive variables. Using a β_0 implies an expected value $E(\varepsilon)=0$, as any non-zero expectation for ε would be absorbed in β_0. By restricting f to this form, the modeling problem is reduced to the estimation of $k+1$ variables, which is a much simpler task. Notice that while the individual parameters β must enter linearly, the predicting variables as well as the observed variable do not have to be linear but can be of any form, *e.g.* log-transformed, which makes linear models a powerful tool for regression analysis even of not strictly linear systems.

A different notation for linear models is the matrix form

$$y = X\beta + \varepsilon \quad ,$$

where $Y = (y_1, y_2, \ldots, y_n)^T$, $\beta = (\beta_0, \beta_1, \beta_2 \ldots, \beta_k)^T$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \vdots & x_{1k} \\ 1 & x_{21} & x_{22} & \vdots & x_{2k} \\ \ldots & \ldots & \ldots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

with *n* observations and *k* predictive variables. The column of ones incorporates the intercept term. As an example, the most simple linear model $y = \beta_o + \varepsilon$, which predicts *y* solely by its mean (often written as $y = \mu + \varepsilon$), can be rewritten as

$$\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix} \beta_o + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{pmatrix} \quad .$$

In most cases the response variable is assumed to be continuous, but the predictive variables can either be continuous, discrete or even categorical. In case of continuous or discrete predictive variables, the model is called a linear regression model while in the case of categorical variables it is called an *analysis of variance* (ANOVA). Depending on the numbers of variables in an ANOVA, it is called a one-way, two-way or multiple-way ANOVA. The special case of a one-way ANOVA with only two categories for the predictive variable is equivalent to the t-test.

Least Square Estimation

The basic idea of linear regression is to estimate the β in the regression equation $y = X\beta + \varepsilon$ that bests separates the systematic influences $X\beta$ from the random error ε, for example those β that lead to the smallest errors ε. The most popular approach to estimate β, which was also utilized in this study, is the least square estimation. It minimizes the sum of the squared errors

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon \quad .$$

Inserting the initial regression equation into this sum leads to

$$\varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta) = y^T y - 2\beta X^T y + \beta^T X^T X\beta \quad .$$

To find a β that minimizes this function the equation is differentiated in respect to β and set to zero. Accordingly, it is the aim to find an estimator $\hat{\beta}$ that satisfies

$$X^T X\hat{\beta} = X^T y \;\rightarrow\; \hat{\beta} = \frac{X^T y}{X^T X} \quad .$$

These equations, which are often called *normal equations*, can be solved using several techniques with the most popular being the 'QR factorization'; however, these topics are beyond the scope of this thesis (the interested reader might refer *e.g.* to Björck[111]). Given an estimated $\hat{\beta}$ the response variable $\hat{y}$ and the residuals $\hat{\varepsilon}$ can then be predicted as

$$\hat{y} = X\hat{\beta} \quad \text{and} \quad \hat{\varepsilon} = y - X\hat{\beta} = y - \hat{y} \quad .$$

Assuming that these residuals are independent and normally distributed it can be shown that the least square estimated $\hat{\beta}$ is the maximum likelihood estimator.[112] Plots of the distribution of the resulting residuals are helpful to check these assumptions visually.

Diagnostic Plots

One of the most important diagnostic plots for a linear model is the *residual vs. fitted* plot. It plots $\hat{\varepsilon}$ against $\hat{y}$ and should result in a symmetric vertical distribution around zero. Figure 10 A-C shows an example of a plot of a valid simple linear regression model (A) and two non-valid models (B and C), where B illustrates a variance dependency of the residuals on the fitted value and C illustrates a nonlinear dependency.

Further, a normal distribution of the residuals must be given to assure $\hat{\beta}$ to be the best possible estimator. To visually inspect this assumption the *quantile-quantile* or *Q-Q*-plot of the residuals against the normal distribution can be used. A common way to generate this plot is sorting the individual residuals $\hat{\varepsilon}_1 ... \hat{\varepsilon}_n$ and then computing

$$q(i) = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

for all $i \in \{1,...,n\}$, where Φ^{-1} is the quantile function of the standard normal distribution. Plotting the sorted $\hat{\varepsilon}_i$ against the computed $q(i)$ should result in a straight line if the residuals are normally distributed. Figure 10 D-F shows three example Q-Q-plots, one of a valid model (D) and two of non-valid models (E and F). The consequence of finding (severe) non-normality can be that the least square estimate might be non-optimal and other estimation techniques should be used preferable like the generalized least square approach or the ridge regression.

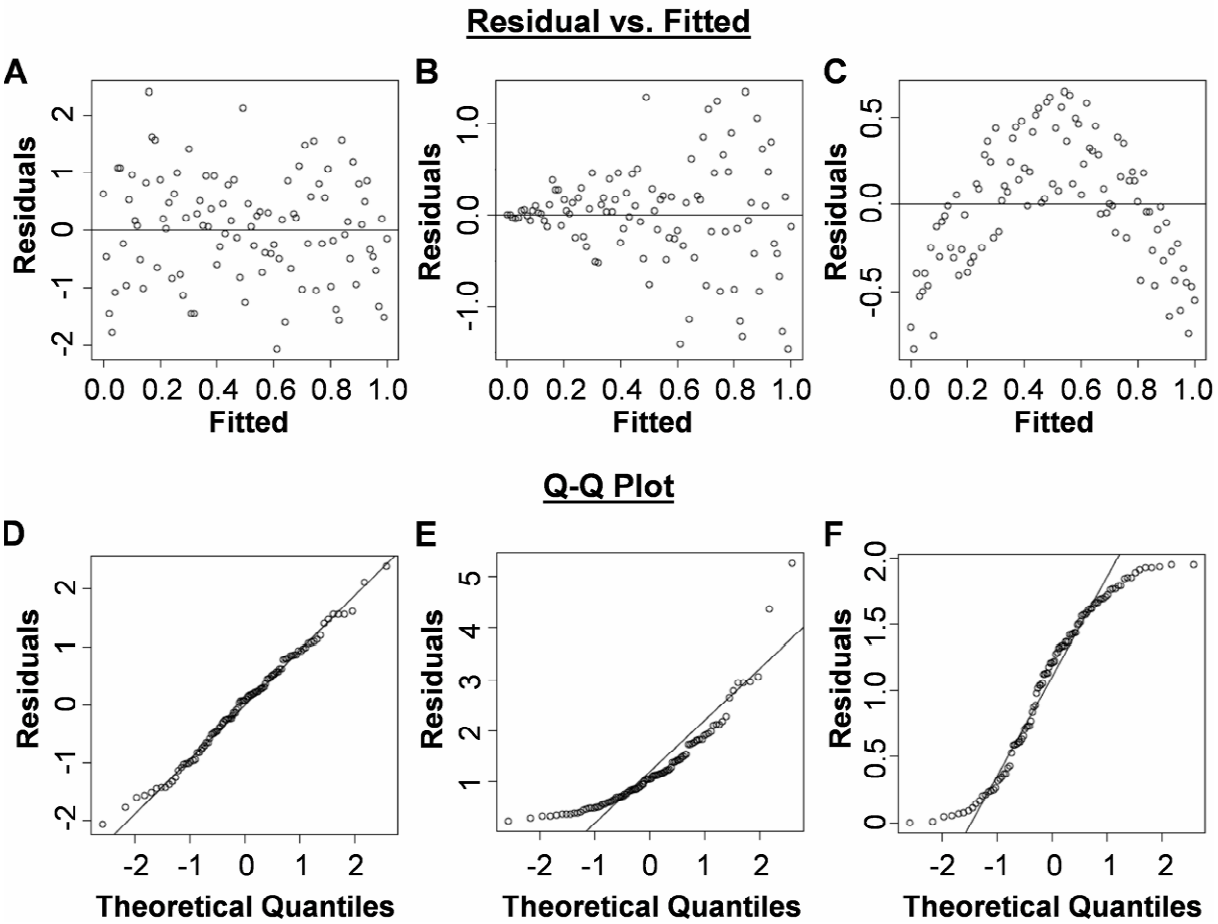

Figure 10: Diagnostic Plots for Linear Models
(A-C) Residuals vs. fitted values plot. **(A)** Valid linear model. **(B)** The size of the residuals depends on the fitted values. **(C)** Non-linear dependency. **(D-E)** Quantile-Quantile-plots of the residuals against the normal distribution. **(D)** Valid linear model. **(E)** Exponentially distributed residuals. **(F)** Uniform distributed residuals.

Goodness of Fit and Significance Testing for Linear Models

To compare different linear models predicting the same observed variable like done in the analysis of correlated enrichment changes in mice hearts, it was of interest how good each individual model fits the observed data. In this study, the diagnostic measure R^2 was used, which is also often called *coefficient of determination* or *percentage of variance explained*. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad ,$$

where $\bar{y}$ is the arithmetic mean of y. It ranges between zero and one, with higher values indicating a better model fit. While R^2 is a good measure to compare different models an overall minimal value for R^2 which separates well from poor models is often hard to establish as different problems and dependent variables can lead to quite different expected ranges of R^2.

Further, hypothesis testing can be used to access the significance of a given linear model. In linear regression model and ANOVA analyses, two hypothesis tests are usually performed. The first is a

simultaneous test for the significance of any predictive variables and the second is the test for the significance of a single predictive variable.

To test the significance of a full model "Ω : $y = X\beta + \varepsilon$" it is compared against the simplest model "ω : $y = \beta_o + \varepsilon$", which describes the response only in terms of its mean. To test if any predictive variable is useful to explain the observed variable the null hypothesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

is formulated. This null hypothesis is tested using the differences in the individual sum of squarred errors combined using the *F* statistic[112]

$$F = \frac{\left(\hat{\varepsilon}_\omega{}^T \hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega{}^T \hat{\varepsilon}_\Omega\right)/(k-1)}{\hat{\varepsilon}_\Omega{}^T \hat{\varepsilon}_\Omega/(n-k)} \quad ,$$

where $\hat{\varepsilon}_\Omega{}^T \hat{\varepsilon}_\Omega$ is the sum of the squared errors for the full model and $\hat{\varepsilon}_\omega{}^T \hat{\varepsilon}_\omega$ is the sum of the squared errors for the simplest model, which is equivalent to $(y-\bar{y})^T (y-\bar{y})$. The F value $F_{k-1,n-1}$ then provides the p-value for the test. If the null hypothesis can be rejected the model likely contains valid predictors for the response variable.

Furthermore, it can be tested for each single predictive variable X_i if it has a significant influence on the overall prediction of *y* given it coefficient β_i. The appropriate null hypothesis is

$$H_0 : \beta_i = 0 \quad .$$

Again the full model Ω is given as above. In addition, the model "Ω-β_i" is defined which is Ω with all its predictive variables except β_i. To test H_0 the same formula to calculate the F value as above is used now exchanging $\hat{\varepsilon}_\omega{}^T \hat{\varepsilon}_\omega$ with $\hat{\varepsilon}_{\Omega-\beta_i}{}^T \hat{\varepsilon}_{\Omega-\beta_i}$ and slight changes for the degrees of freedom.

2.3.2 Normalization of Large-scale Data

In this study, a large number of high-throughput datasets like ChIP-chip or siRNA knockdown experiments were analyzed. However, these datasets contain systematic variations and biases that are inherent to the experimental process. Sources of these variations are different initial amounts of DNA/RNA that were used in each experiment, different efficiencies of the reverse transcription (for RNA), different efficiencies in the labeling or the detection process and so on. If not accounted for correctly, these biases will significantly impact on the quality of the analysis and can drive any results invaluable. Therefore, a key factor in a reasonable bioinformatic analysis is the selection of proper methods to correct these systematic effects. This process is called normalization.

In this thesis, two advanced but inherently different technique to normalize the high-throughput data were selected and applied. The first is the variance stabilization normalization (vsn) method for

microarray data introduced by Huber *et al.*[113] The second is the qspline normalization approach by Workman *et al.*[114] Both methods were already successfully applied to normalize high-throughput datasets.[115-117] The two methods differ as qspline is a non-parametric normalization (does not make any assumption on the sort of appropriate transformation), while vsn is a parametric normalization. Vsn was used as the standard normalization technique in this thesis (*e.g.* for the normalization of the ChIP-chip data), as a parametric solution should in general perform better if its assumptions hold. However, as found for the siRNA expression data (section 3.1.1), it can lead to worse results if the assumptions do not hold. In this case it was replaced by the non-parametric qspline normalization.

Simple Normalization

For the following sections a matrix Y is assumed containing measurements of K probes from N experiments. For data normalization in general, two tasks arise: the intra- and the inter-experimental normalization. The first, which is often named correction instead of normalization, is the task of finding and removing trends that are inherent to the data of a single experiment N_i. An example is the mean-variance dependency, which will be discussed below. Further, inter-experiment normalization refers to the task of manipulating data to make measurements from different experiments comparable. A common assumption in normalization methods is that most of the measured data are the same between the different experiments and therefore overall attributes of the data distributions like quantiles can be used to normalize experimental variations. Common graphical tools to assess the similarity of distributions before and after normalization are scatterplots, which are only useful to compare two experiments, and boxplots and density plots, which can be used for larger numbers of experiments.

The simplest method to normalize a single measurement $y_{k,n}$ is to calculate the mean intensity and range (difference between the smallest and the biggest value) of every experiment $n \in \{1,...,N\}$, define a "target" mean and range *e.g.* using the average of the calculated means/ranges or arbitrarily using a mean of 0 and a range of 1 and apply an affine linear transformation of each value in each experiment like in

$$y_{k,n}^{adj} = \alpha_n + y_{k,n}\lambda_n \quad ,$$

where $y_{k,n}^{adj}$ is the resulting adjusted value, α_n is the linear shift that is needed to adjust the mean intensity of the experiment n to the target distribution and λ_n is the experiment specific scaling parameter to adjust the range. However, such a simple transformation will in most cases only be valid for very simple datasets. The main reason is that the systematic errors that lead to differences in experiments are not linear but their relation is dependent on the signal intensity. A common way to visualize this dependency is to plot the differences between the logarithmic intensities versus their sum or mean. An example for such a plot, which is often called MA-plot (M for *minus* and A for *add*), is given in Figure 11 A which was taken from the analysis of the siRNA data performed in this study. It is easily seen that the differences between the two experiments is dependent on the signal intensity, often resulting in 'banana'-shaped plots. Such a dependency, if not normalized for, will give rise to different powers to detect differences in the experiments, which will be dependent on the individual

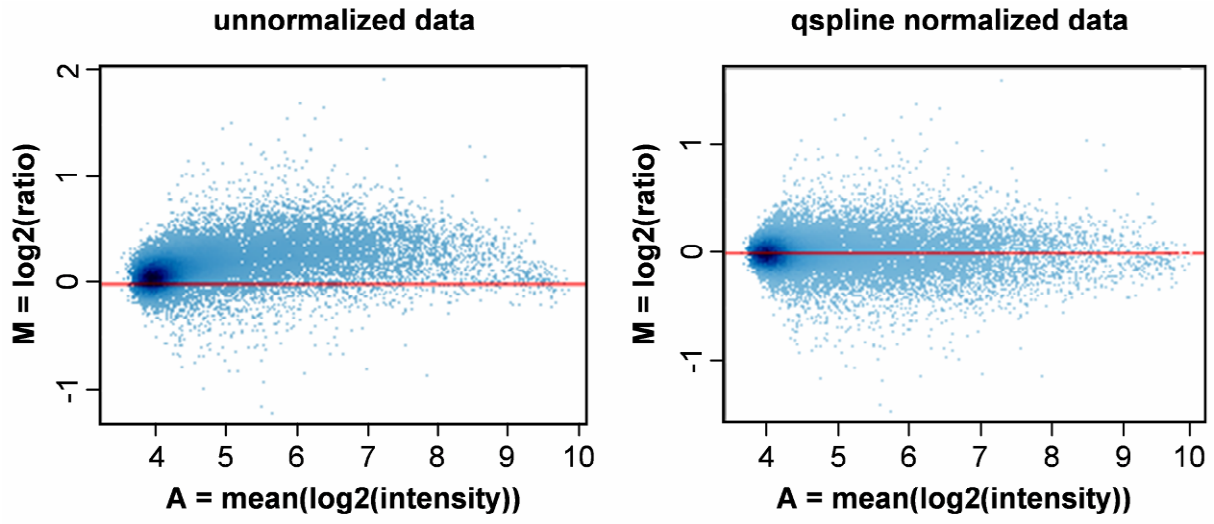

Figure 11: Example MA-plot and qspline normalization taken from the siRNA analysis
Example of an MA-plot **(A)** before and **(B)** after qspline normalization. The example is taken from the Gata4 siRNA expression data normalization (section 3.1.1).

signal intensity of the probe under study. Therefore, large-scale datasets like microarray experiment require more sophisticated normalization approaches like the used vsn and qspline normalization explained below.

Which normalization method to choose for a certain series of experiments is a question that is often difficult to answer. In general, one should aim at using the normalization method that leads to the most accurately measures with the lowest bias, *e.g.* which leads to the "correct" set of differentially expressed genes. However, in this study and in most cases this correct set is (at least partially) unknown and can therefore not be used to determine the right normalization method. Instead, a further measure of the quality of normalization, the similarity of replicated experiments, was used. Replicated experiments are a valuable selection aid as they should in general only reflect the experimental variations and no real differences.

Variance stabilization normalization

The variance stabilization normalization by Huber *et al.* is based on a specific model of the relationship between the differences or variance of several experiments and their signal intensity. This dependency can better be inspected visually for two experiments in a variation of the previously introduced MA-plot which shows the difference (M) between measurements versus the rank of their sum (A_{rank}). Figure 12 A shows an example based on one of the ChIP-chip experiments analyzed in this study. It is easily seen that the variance increases with the intensity. Huber *et al.*[113] model this dependency based on the model for measurement error of gene expression from Rocke and Durbin[118] and divide it into an additive and a multiplicative component assuming the quadratic mean-variance dependency

$$Var_{y_k}\left(\bar{y}_k\right) = \left(c_1\bar{y}_k + c_2\right)^2 + c_3 \quad ,$$

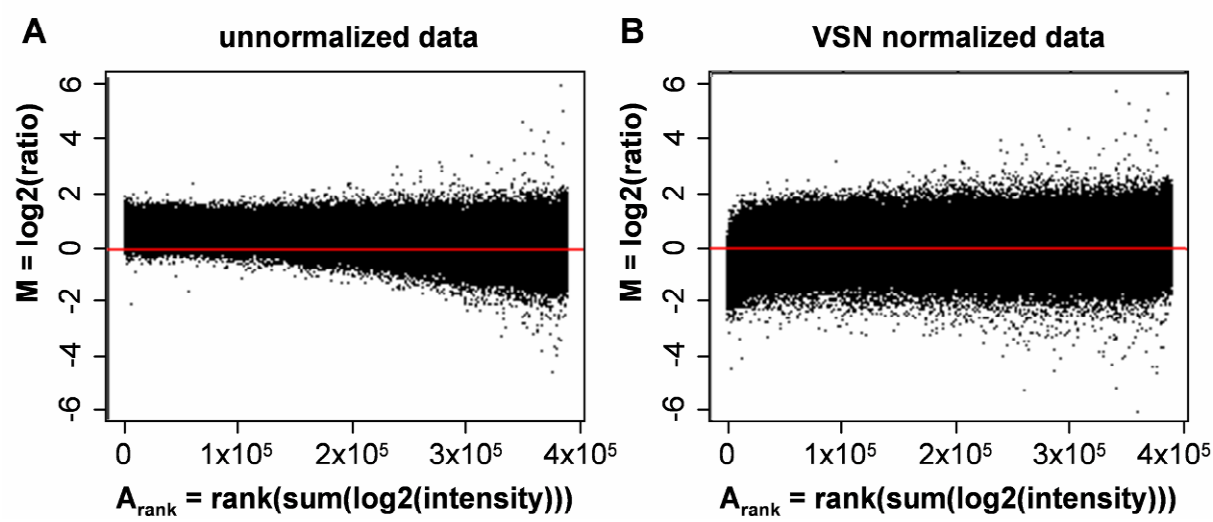

Figure 12: Example of MA_{rank}-plot and vsn normalization taken from the ChIP-chip analysis
Example of an MA_{rank}-plot **(A)** before and **(B)** after vsn normalization. The example is taken from the Srf ChIP-chip normalization (section 3.1.1).

where Var_{y_k} is the variance of gene k over all experiments n, $\bar{y}_k$ is its mean and the three parameters c_1, c_2 and c_3 parameterize the assumed dependency. Beside the motivation given by Rocke and Durbin's model this form of dependency was shown to be present in many microarray experiments.[113] To make the variance independent of the mean, the general asymptotic variance stabilization transformation[119]

$$h(y) = \int^{y} \frac{1}{\sqrt{Var_y(\bar{y})}}\, d\bar{y}$$

is used. Incorporating the previous mean-variance dependency based on the additive and multiplicative error model into this variance stabilization transformation, Huber *et al.* construct an areasinus hyperbolicus (arsinh) transformation of the data to stabilize the variance:

$$h(y) = \gamma \operatorname{arsinh}(a + by) \quad .$$

They relate this formula to the above additive/multiplicative error model by

$$\gamma = \frac{1}{c_1},\ a = \frac{c_2}{\sqrt{c_3}} \text{ and } b = \frac{c_1}{\sqrt{c_2}} \quad .$$

Huber *et al.* interpret this arsinh function as a form of generalized logarithm which is also valid for negative values. However, the variance stabilization is *per se* no method to normalize systematic effects between different experiments but only addresses the dependency of the variance on the mean intensity in each single experiment. Therefore, the vsn method further incorporates an affine linear normalization as described above, gaining

$$h_n\left(y_{k,n}^{adj}\right) = \operatorname{arsinh}\left(a + b\left(\alpha_n + y_{k,n}\lambda_n\right)\right) = \operatorname{arsinh}\left(a_n + b_n y_{k,n}\right)$$

with $a_n = a + b\alpha_n$ and $b_n = b\lambda_n$. Note that the overall scaling factor γ was omitted. Finally, the parameters a_n and b_n of this variance stabilized normalization needs to be estimated for every experiment. To do so, Huber *et al.* reformulate their normalization in terms of the expected independency of the variance from the mean using the linear model

$$h_n\left(y_{k,n}^{adj}\right) = \beta_{0,k} + \varepsilon_{k,n} = \mu_k + \varepsilon_{k,n} \quad ,$$

where $\beta_{0,k} = \mu_k$ again is the mean of $h_n\left(y_{k,n}^{adj}\right)$ and $\varepsilon_{k,n}$ is the experiment and probe-specific error term which should have a constant mean equal to zero and a variance that is equal to the common variance. The parameters are then estimated in an iterative way alternating between a least square fit based on a subset $\hat{k} \in \hat{K} \subset \{1,...,K\}$ and choosing $\hat{K}$ as the set of probes with the smallest residues. By using only a subset of all probes for the estimation, the method accounts for probes that might reflect real differences and should therefore be excluded from the estimation of the normalization parameters (for a detailed description of the estimation process refer to Huber *et al.* 2003[120]). Transforming the data of the different experiments using the estimated parameters and the arsinh function will then result in a common mean and scale of all experiments with a variance that is independent from the mean as illustrated in Figure 12 B.

Qspline Normalization

Unlike vsn, the qspline normalization method by Workman *et al.* is a non-parametric normalization technique which tries to fit the distributions of several experiments to a common "*target*" distribution. This common distribution is defined by taking the arithmetic mean $\bar{y}_k$ of each probe over all N experiments. Then, for every experiment n a number of q quantiles points are chosen from this common distribution and from the experiment at the same time and a natural cubic spline function is interpolated between the experiment and the common distribution using the sampled q points. A spline function is a curve fitting function that is defined in a piecewise manner, using polynomials to fit the data between two successive points. Special requirements are put on the crossover region between one polynomial and the next to ensure smoothness. The resulting spline function is then used to transform all probe's measurements in experiment n to the common distribution. To gain more robust results, the spline function interpolation is repeated r times with increasing equidistant offsets from the original q quantile points. Finally, the normalized value for a probe k is calculated by the mean over its transformed values in all repeats. In this analysis, q was set to 100 and r was set to 5 as suggested in the original publications by Workman *et al.*[114] Applying qspline normalization to a set of experiments will center any systematic difference between two experiments on zero. This effect is shown in Figure 11 B. Therefore, qspline normalization is similar to quantile normalization as both lead to the same distribution in all normalized datasets, however qspline makes more relaxed assumptions on the data in general and will maintain the individual effects of each experiment.[115]

2.3.3 Pairwise Distance Measures and Clustering

Given large datasets comprising multiple entities (*e.g.* genes and expression data) it is often desirable to measure the amount of dependency between these. Different techniques to determine this dependency have been utilized in this study, *e.g.* to find correlated ChIP enrichment changes in the analysis of time series ChIP-qPCR data (section 3.2.2) or to determine genes with highly similar expression patterns as performed in the patient analysis (section 3.3.5).

In the following, a set of entities *g* is given that have been measured in a number of experiments *c*. A measurement of a specific entity in a specific experiments will be called x_i, where $x \in g$ and $i \in c$. In case of gene expression data, *g* refers to the set of genes and *c* could refer to a number of conditions, individuals or time-points. The dependency between two entities *x* and *y* is then measured by calculating a distance metric incorporating the individual measurements x_i and y_i.

Common Pairwise Distance Measures

A simple and widely used distance measure is the Euclidean distance

$$d(x,y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

between the two entities *x* and *y*. However, throughout this study, the Pearson correlation coefficient

$$r_{x,y} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

where $\bar{x}, \bar{y}$ again are the arithmetic means over all x_i and y_i, will be used, which is more popular especially for gene expression data. Spearman's rank correlation coefficient, which replaces the individual values by their ranks, is the non-parametric alternative to the Pearson correlation coefficient. Both measure the linear statistical dependence between the two variables and range between 1 (correlation) and -1 (anti-correlation) with a zero value implicating no linear dependency. They were shown to have high power in the analysis of expression data in a number of species under a steady-state.[121-123] However, both distance metrics have a major drawback as they measure only linear relations between the two expression vectors. This poses no threat as long as only linear relations are of importance. However, for its use in the analysis of gene expression data it was demonstrated that the Pearson correlation coefficient can be distorted if the expression levels show a non-uniform distribution across the expression patterns.[124]

Mutual Information

The information theoretic concept of mutual information is a reasonable way to overcome the mentioned limitation of linearity As the prediction of genes with common regulators performed in the

analysis of gene expression in patient with congenital heart disease (section 3.3.6) was highly dependent on the accuracy of the implemented distance measure, mutual information was additionally integrated to determine the existence of non-linear dependencies in gene expression data. Mutual information was introduced to gene expression analysis to provide a more general measure of dependencies in the data, in particular, positive, negative and nonlinear correlations.[125] It is a well known measure in information theory[126] that has been used to analyze gene-expression data in a number of studies.[124,127-129] The definition of mutual information is based on the Shannon entropy, which is defined as

$$H(A) = -\sum_{i=1}^{N} P(a_i) \log P(a_i) \quad ,$$

where A is a system with N possible states yielding any of the possible values a_i with the probability $P(a_i)$.[126] The joint entropy $H(A,B)$ of two discrete systems A and B is defined analogously

$$H(A,B) := -\sum_{i=1}^{N} \sum_{j=1}^{M} P(a_i, b_j) \log P(a_i, b_j) \quad ,$$

where $P(a_i,b_i)$ denotes the joint probability that system A is in state a_i and system B in state b_i. If the two systems are statistically independent the joint entropy becomes

$$H(A,B) = H(A) + H(B) \quad ,$$

which leads to the definition of the mutual information[127]

$$I(A,B) := H(A) + H(B) - H(A,B) \quad \geq 0 \quad .$$

The mutual information is strictly positive and becomes zero if no dependency between the two systems exists. Using the entropy definition given by Kullback,[130] this formula can be rewritten as

$$I(A,B) = \sum_{i=1}^{N} \sum_{j=1}^{M} P(a_i, b_j) \log \frac{P(a_i, b_j)}{P(a_i) P(b_j)} \quad .$$

This definition requires explicit knowledge of the respective probability distributions. However, in most cases these are unknown and have to be estimated. The most straightforward and widely used approach is to utilize a histogram based technique.[124,128] An alternative method, which was proposed by Moon *et al.*[131] is based on kernel density estimation. As it was found to be superior to the histogram methods[131] it was therefore applied in this study. The method by Moon *et al.* uses the Gaussian kernel

$$f(x) = \frac{1}{N}\frac{1}{h\sqrt{2\pi}}\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2h^2}\right)$$

to estimate the one-dimensional probability distributions and

$$f(x,y) = \frac{1}{Nh^2}\frac{1}{2\pi}\sum_{i=1}^{N}\exp\left(-\frac{\left(\sqrt{(x-x_i)^2+(y-y_i)^2}\right)^2}{2h^2}\right)$$

to estimate the two-dimensional distribution. Then the mutual information is computed as

$$I(X,Y) = \int_x\int_y f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy$$

using an optimal bandwidths *h.* Following the argumentation of Steuer *et al.*[127] and assuming that the analyzed gene expression measurements are a faithful sample of the underlying probability distribution this was further simplified to

$$I(X,Y) = \frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{f(x_i,y_i)}{f(x_i)f(y_i)}\right] \quad .$$

Mutual information has been used in a number of studies to infer undirected gene networks often coupled to a fixed threshold.[124,132] The concept was further enhanced by removing indirect edges between genes that were part of a regulatory chain by Margolin *et al.*[133,134] using the data-processing inequality principle. This principle states that, if a gene *x* regulates another gene *y* via a gene *z*, the mutual information between *x* and *z* is smaller than between either *x* and *y* and *y* and *z*. While this was shown to successfully reconstruct genetic networks[135] it provides the possibility to miss important interactions, if *e.g.* gene *x* is indeed directly regulating gene *z* or if feed-forward loops, which are important motifs in regulatory networks, exist. Another approach to extend the power of mutual information which was not used in this study but has been applied to reverse engineer genetic networks is the use of conditional mutual information,[136,137] where the mutual information between gene *x* and *y* is calculated conditioned on gene *z*.

Odds Ratios

Correlation coefficients and mutual information are often used when the statistical dependency between two continuously variables like gene expression should be computed. In the easier case of two binary variables, another measure can be used, namely the so-called '*odds ratio*'. In this study, odds ratios have been used to determine the statistical dependencies between target genes of TFs in ChIP-chip (section 3.1.6) or differentially expressed genes found in the analysis of siRNA data

(section 3.1.8). For two binary variables A and B with the probabilities $P(A)$ and $P(B)$ and their complementary probabilities $1-P(A)$ and $1-P(B)$ their odds ratio is defined as

$$\text{Odds Ratio}(A,B) = \frac{\frac{P(A)}{1-P(A)}}{\frac{P(B)}{1-P(B)}} = \frac{P(A)\cdot(1-P(B))}{P(B)\cdot(1-P(A))} \quad .$$

Odds ratios can be easily derived from a contingency table *T*

	$P(A)$	$1-P(A)$
$P(B)$	a_1b_1	a_0b_1
$1-P(B)$	a_1b_0	a_0b_0

as

$$\text{Odds Ratio}(T) = \frac{a_1b_1 \cdot a_0b_0}{a_1b_0 \cdot a_0b_1} \quad .$$

They range between 0 and ∞ and are a measure of how more likely $P(A)$ is given *B* and vice versa.

Clustering

To find co-regulated genes and predict binding sites of common regulators in human, the aforementioned Pearson correlation coefficient was implemented and compared to mutual information (section 3.3.5 and 3.3.6). However, the calculation of a distance metric is often insufficient to infer co-regulatory dependencies. The main problem is to define thresholds that separate the distance matrix entries into those values that infer regulatory dependencies and those that only reflect background. A simple and widely used method to fulfill this task is clustering. Given a matrix of pairwise distances, clustering arranges entities into groups with similar measurement profiles.

In this study, the method selected to cluster gene expression and phenotype annotation vectors was hierarchical clustering which is one of the most popular clustering techniques due to the simplicity of its concept and no requirement to define the number of cluster *a priori*. In every step the two most similar entities are joined into one entity, the distances between this new entity and all other entities are recalculated according to a fixed algorithm (*e.g.* arithmetic mean or maximum) and the clustering progresses to the next cycle until all entities are finally joined. The sequence of joined entities is reflected in the cluster dendrogram of distances between the individual sub-entities. If needed, a threshold can be used to cut the dendrogram at a specific height producing groups of genes with similar expression patterns. Another widely used clustering algorithm, which was not implemented in this study, is *k-means*[138] which partitions the genes into *k* clusters according to their nearest center. Instead of choosing a threshold *k-means* requires the definition of the number of cluster *k a priori* which can have a very crucial influence on the correctness of the results. More advanced clustering

techniques like the self-organizing maps[139] or self-organizing tree[140] algorithms have been proposed that successfully circumvent the selection of a fixed number of clusters. In addition, several clustering algorithms have been proposed that do not require all genes to show a high similarity under all conditions or time points but only under a subset of these. This technique is usually called bi-clustering and has been introduced to expression data analysis by Cheng and Church.[141] Numerous bi-clustering algorithms exist,[142-145] however, they all require a large number of measurements under different conditions or time points which is often not given in biological experiments for higher vertebrates.

2.3.4 Correction for Multiple Testing

Analyzing large-scale biological data like the ChIP-chip or siRNA expression analysis performed in this study involves the repeated performance of statistical tests. The main problem of this multiple hypothesis testing is an accumulation of the false positive rates of the individual tests which will lead to an overall higher chance of falsely rejecting at least one tested null hypothesis thereby increasing the chance of false positive discoveries.

Classical methods to correct for this increase in false positives have tried to ensure a least overall significance level by adjusting the individual hypothesis significance levels. The most straightforward method proposed was the Bonferroni correction which distributes the overall significance level α evenly on all conducted tests by requiring a significance level of at least α/n, where n is the number of tests. However, this method and other methods that try to ensure a least overall significance level are too conservative especially for the analysis of high-throughput data where the number of tests can easily exceed many thousands.[146] Applying these classical methods will result in only very low numbers of significant tests. Instead, methods that control the false discovery rate (FDR), which is the expected proportion of false discoveries among all significant tests, have been introduced for these kinds of analyses. The method to control the FDR in this study was originally introduced by Benjamini and Hochberg[147] for independent p-values in 1995 and was later adapted by Benjamini and Yekutieli[148] in 2001.

To ensure that an expected FDR is less than a given δ both the Benjamini-Hochberg and the Benjamini-Yekutieli method sorts the $P_1...P_m$ p-values resulting from m different hypothesis tests in increasing order and then find the largest index $k \in i$ where

$$P_i \leq \frac{i}{m \cdot c(m)} \delta \quad .$$

Subsequently, all the hypothesis with p-values smaller or equal to P_k are rejected. The difference between the two methods lies in the definition of $c(m)$. While the original Benjamini-Hochberg method used $c(m)=1$, Benjamini and Yekutieli showed that this is only valid for independent p-values. Instead they proposed

$$c(m) = \sum_{j=1}^{m} \frac{1}{i} \quad ,$$

which results in more conservative estimations of the FDR but does not require independency of the p-values (refer to Benjamini and Hochberg[147] and Benjamini and Yekutieli (2001)[148] for a detailed discussion). As this independency of p-values cannot be guaranteed for the analysis of gene transcription networks in general the more conservative approach was used throughout this study.

Finally, Benjamini-Yekutieli FDR-adjusted p-values P^{adj} can be computed using the step-wise procedure

$$P_i^{adj} = \begin{cases} P_i & \text{for } i = m. \\ \min\left(P_{i+1}^{adj}, \frac{m \cdot c(m)}{i} P_i \right) & \text{for } i = m-1, m-2, ..., 1. \end{cases}$$

The FDR-adjusted p-value thereby represents the lowest level of FDR, where the appropriate hypothesis belongs to the set of rejected hypothesis for the first time.[149-151]

2.3.5 Analysis of ChIP-chip Data

To determine binding sites of cardiac regulators, ChIP-chip experiments of several TFs have been conducted. However, different from *e.g.* qPCR gene expression data, ChIP-chip data cannot be interpreted in a straight-forward manner but requires a sequence of individual steps to make the results interpretable. Most importantly, individual probes that reflect a significant enrichment between the ChIPed and the Input measurement must be detected and combined to determine binding sites. The pipeline implemented for ChIP-chip peak detection in this study was originally developed in collaboration with the group of Dr. Wolfgang Huber of the EMBL in Heidelberg in an analysis of histone modification ChIP experiments[16] and was later implemented in the R[109] package 'Ringo' by Toedling *et al.*[152] The pipeline is divided into two steps: the first is the normalization of the array data and the second is the actual peak calling.

ChIP-chip data normalization

As ChIP-chip is based on arrays, normalization of probe intensities is a main issue due to non-specific binding of DNA fragments to the probes on the array. If not accounted for, the resulting biases can severely deteriorate any subsequent analysis.[153] Due to the high similarity, a number of normalization techniques have been adapted from gene expression analysis. These include the quantile-based and vsn normalizations which have been described in section 2.3.2.

In this study, the method used to normalize the ChIP-chip array intensities of each channel was *vsn* as visual inspection of MA_{rank}-plots revealed a clear variance dependency. After normalization, log-ratio enrichment levels for each probe were subsequently calculated by subtraction of log Cy3 (Input) from log Cy5 (ChIP sample) for every probe and every ChIP pool.

As can be seen in Figure 13, normalized ChIP enrichments vary greatly between nearby probes and even between replicates, possibly due to the effect that different probes measure the same target DNA amount with different efficiencies, caused by different qualities of probe synthesis on the array, probe GC content, target cDNA secondary structure, cross-hybridization or other reasons. To cope with this

variability, a smoothing approach was used that places a window of size 600 bp (the estimated fragment size) centered over each probe position p_i and exchanges the original probe level by the median of all probe levels (called z_i) that falls into that window, thereby summarizing multiple measurements per probe through integrating both pools into one value. The resulting "smoothed" running mean is shown as a yellow line in Figure 13.

ChIP-chip peak calling

Several methods have been proposed to detect enriched binding sites from normalized ChIP-chip data. In general, they divide into two categories: some require a number of enriched probes inside a local area while the other assume a specific shape of probe enrichments. The method implemented for ChIP-chip peak calling in this study belongs to the first category. To call a specific probe position *enriched* it was required to have a smoothed probe level that is greater than a threshold t_0. In line with the finding of Buck *et al.*[154] that the signals from ChIP-chip experiments are heavily right-tailed with the left tail of the distribution very likely resulting from background noise, the threshold was chosen according to a null distribution of the smoothed probe levels as follows: the mode m of the distribution was calculated and all scores that are lower than m were taken twice, once with their original value z_i and once as $m + (m - z_i)$ thereby "mirroring" the distribution from the left side of the mode to the right side. The idea behind this approach is to estimate a random distribution of not enriched probes from the probes with the weakest signal. The empirical cumulative distribution function of these mirrored z-scores then allows to obtain a p-value for comparing each smoothed probe level's height to the null distribution. These P-values were corrected for multiple testing using the described *Benjamini–Yekutieli* FDR approach[150] and probe levels with an FDR smaller than 0.1 were called significantly enriched. Significant probe positions that were less than 200 bp (twice the tiling) apart from each other are finally combined and regions containing at least 3 such probes are called *peaks*.

Alternative approaches that also belong to the first category of ChIP-chip peak calling have been proposed by Keles *et al.* [155] and Cawley *et al.*,[156] which applied t or Wilcoxon rank sum test statistics, respectively, to sliding window approaches to find enriched regions. Further, Li *et al.*[157] and Ji *et al.*[158] introduced hidden Markov models to find locally enriched region. An approach of the second category has been implemented by Zheng *et al.*[159] Based on a direct modeling of the DNA fragmentation process they assumed that relevant peaks should have a triangular shape when the ChIP-chip signal is transformed to log scale. Two recent approaches are hard to classify into one of the two categories. Keles[160] and Gottardo *et al.*[161] each use Bayesian hierarchical models to identify TFBS from ChIP-chip data. While they both use the superior Bayesian statistical framework, they pose some extra requirements on the ChIP-chip data, like multiple replicated experiments or a very sparse distribution of peaks, which are not easily fulfilled.[162] The peak finding method implemented in this study was found to be superior to other implemented methods at the time of the analysis as it was able to successively recover a predefined set of 42 known target genes for the respective TFs.

2.3.6 Analysis of ChIP-seq Data

In addition to the ChIP-chip data, genome-wide ChIP-seq experiments have been analyzed in this study. While ChIP-seq in general has many similarities to ChIP-chip the bioinformatic challenges are different. The task of analysing ChIP-seq data is divided into three steps: the first is the mapping of the

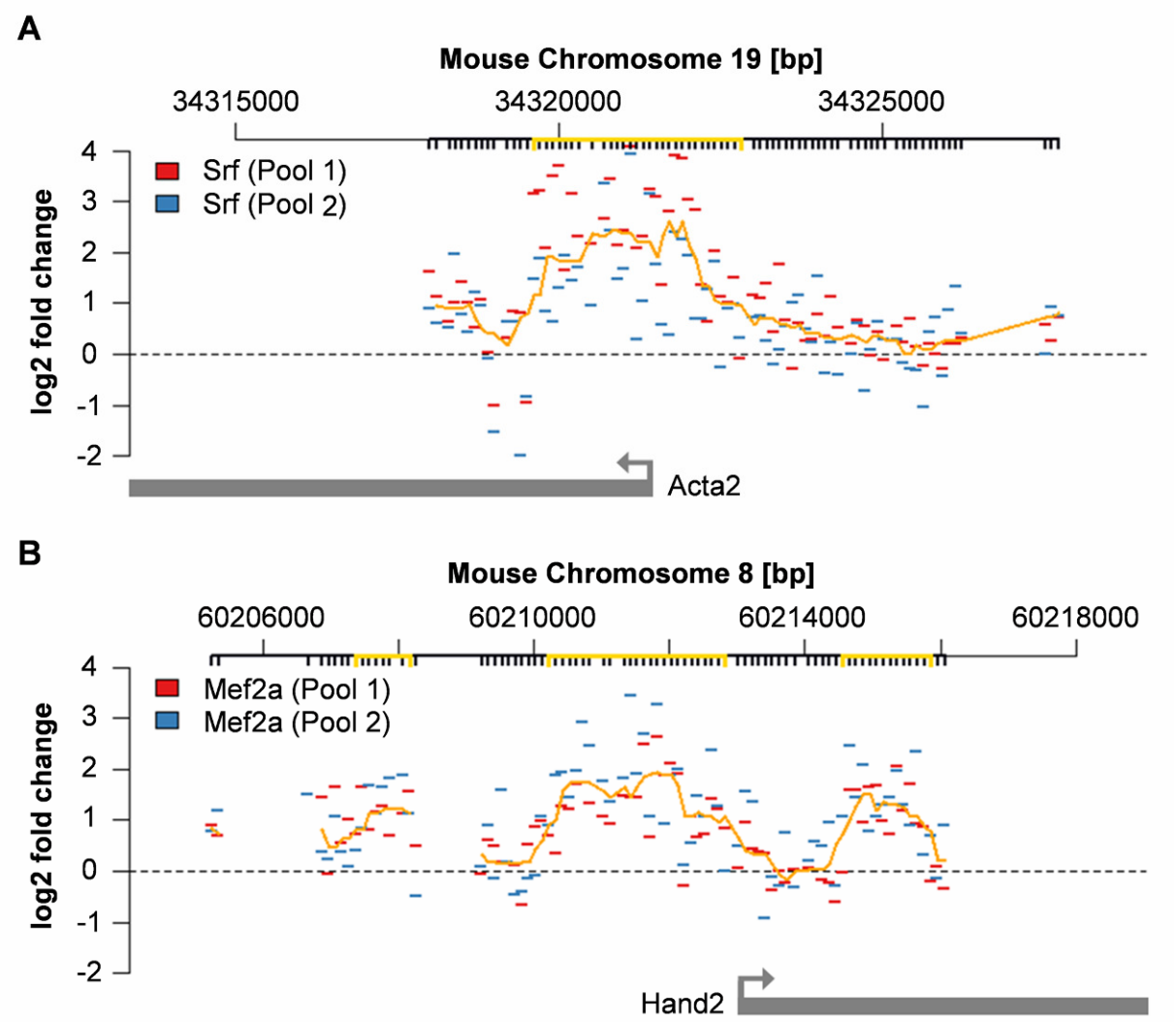

Figure 13: Two examples of ChIP-chip peak calling
Promoter regions and probe enrichment for the two gene-TF pairs **(A)** Acta2 and Srf and **(B)** Hand2 and Mef2a (gene bodies and transcriptional start sites are indicated with gray boxes). Each tick on the y-axis represents a probe on the array. Measured enrichments are indicated below for pool 1 (red) and pool 2 (blue). The "running" median is indicated by an orange line. Called peak regions are indicated by yellow lines on the y-axis.

obtained reads to the reference genome, the second is a normalization of read counts to account for experimental differences between different sequencing runs and the third is the calling of enriched sites or peaks. As only ChIP-seq experiments using a single sequencing run per experiment were performed for this study, no normalization of the resulting reads must be performed, leaving the two tasks read mapping and peak calling.

Read mapping

For the task of mapping the obtained reads to the reference genome a panel of read mappers exist (*e.g.* Weese *et al.*,[163] Langmead *et al.*[164] and Li *et al.*[165]), which often tolerate the occurrence of a specified number of errors or even gaps in the alignment. Given the large number of reads that result from even a single sequencing experiment, the main basis to choose between the available tools is their sensitivity, meaning the number of mapping reads that are missed (if any), and their running time. In this study, the approach by Weese *et al.*[163] was used which is implemented in the RazerS

mapping tool. RazerS was found to be the most sensitive tool for next-generation sequencing experiments while maintaining very short running times.[163] RazerS supports Hamming and edit distance read mapping with configurable sensitivity. It consists of three parts: the filtration parameter estimation part, the filtration part and the verification part.

The filter parameter estimation part is introduced to find the parameters that optimize the running time while guaranteeing a user-defined minimal sensitivity for the filtering part. RazerS is well suited for the analysis of next-generation read mapping as it directly incorporates the typical position-dependent error profile of sequencing reads, as *e.g.* derived by Dohm *et al.*,[166] which reflects the increasing proportion of errors at the end of the reads.

The filtering part, the step that is most crucial in terms of running time, aims to find positions in the reference genome, which are very likely to contain matches to a given read. RaserS implements a *q-gram counting* strategy using the *q-gram lemma*, which states that two sequences of length s with Hamming distance h share at least

$$t = s + 1 - (h + 1)q$$

common substrings of length q, so-called q-grams or q-hits.[167,168] To find likely matching regions, the reference genome is scanned linearly for regions with a number of at least t q-hits using a precompiled index of overlapping q-grams from all reads.

The verification part finally scans each region that passed the filter by counting the actual number of mismatches between the read and the region (for Hamming distance) or using hardware optimized dynamical programming to calculate the edit distance between the two. A true match is a region which has an edit or Hamming distance less than the user-defined cut-off.

After reads have been aligned to the reference genome, it remains to be decided if only reads that can uniquely be mapped to the genome or also those with a number of possible positions are taken into account. If only uniquely mapped reads are taken into account some true binding sites will be lost because they are located in repetitive or duplicated genomic regions. Conversely, allowing multireads will likely improve some true signals but risks the danger to create false-positives. For the read mapping of next-generation sequencing data throughout the analyses, RazerS was used with the simpler Hamming distance mapping approach allowing two mismatches at most and no indels. In addition, only those reads that could uniquely be mapped to the genome were retained.

ChIP-seq Peak Calling

After read mapping and filtering, peak calling aims to identify regions that show significantly more reads than what would be expected by chance. Therefore, a key component of ChIP-seq peak calling is to understand what level of enrichment is required to distinguish signal from noise. In line with the ChIP-chip algorithms, many ChIP-seq peak calling algorithms are based on a sliding window approach. If a certain window has a number of reads that exceeds a defined significance threshold, then this region is called a peak. Distributions used to call significantly enriched windows are typically the Poisson[169,170] or negative binomial distribution.[94] Some algorithms determine the background distribution from a control experiment if available,[94,169-172] while others model the background solely from the ChIP sample itself.[173,174] A number of algorithms further use the strand specificity of resulting

reads to shift positive and negative strand reads by a common distance either to increase the statistical power of the peak detection[169,171,174] or to reduce the number of false-positive peaks subsequently.[94]

Due to the difference in the experimental data (intensity of probes versus read counts) a different method was applied to call enriched binding sites in ChIP-seq, namely the approach from Ji *et al.*[94] (implemented in the CisGenome package) which is suitable to call ChIP peaks for experiments without input data. Its main advantage is the use of a negative binomial distribution instead of the more frequently used Poisson distribution. Modeling both a Poisson and a negative binomial distribution on ChIP-seq Input data from mouse embryonic stem cells and comparing it to the observed Input data, Ji *et al.* showed that the negative binomial distribution is much better suited than a Poisson distribution to model the background distribution in the absence of Input data. At the time of the analysis, the approach by Ji *et al.* was the only one that incorporated the negative binomial distribution and it was consequently used to call enriched peaks in the ChIP-seq data.

To call peaks, Ji *et al.* use a sliding window approach to count the number of reads u in all non-overlapping windows of a specified length w over the whole genome, yielding a vector $u_0 ... u_{max}$ of counts for every u. Their use of the negative binomial distribution is motivated by the Poisson distribution, which defines the probability of finding a number of u_x reads as

$$P_\lambda(u = u_x) = \frac{\lambda^u e^{-\lambda}}{u_x!} \quad .$$

While the Poisson distribution assumes a constant rate λ in all genomic loci, Ji *et al.* drop this assumption by defining λ itself to be a random variable. They encourage this by found positive correlations between ChIP and Input count data, where windows with higher number of reads in the ChIP sample also show higher number of reads in the Input sample, which contradicts with the assumption of an overall constant rate λ. Instead they assume λ_i, the λ of window i, to be distributed according to the gamma distribution

$$\gamma(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta} \quad ,$$

where $\Gamma(\alpha)$ is the gamma function and $\Gamma(\alpha) = (a-1)!$ for positive integers. The gamma-distribution itself is locus-independent but leads to randomly sampled λ_i for each window i. For positive integer values as in ChIP-seq count data, exchanging a fixed λ with $\gamma(\lambda)$ is equal to the negative binomial distribution

$$P_{NB}(u = u_x) = \binom{u_x + \alpha - 1}{\alpha - 1} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{\beta}{\beta+1}\right)^{u_x} \quad .^{94}$$

To define the background model from the ChIP-seq data the α and β are estimated using counts for windows containing no or only a very small number of reads. The observed numbers u_i are

```
markedGenes = Ø
significantNodes = Ø
dagLevels = get the GO DAG levels list
for i from max(dagLevels) to 1
        for u in nodes(dagLevels, i)
                genes[u] = genes[u]\markedGenes[u]
                significantNodes[u] = HypergeometricTest(genes[u], significantGenes)
                if significantNodes[u] ≤ threshold then
                        for x in ancestors(u)  # all nodes on any path from u to the root
                                markedGenes[x] = markedGenes[x] ∪ genes[u]
                        end
                end
        end
end
return significantNodes
```

Algorithm 1: GO enrichment analysis algorithm according to Alexa *et al.*[175] using a hypergeometric test to find significant overrepresentations as suggested by Falcon *et al.*[176]

compared with the expected numbers according to the null model and the ratio between the two is used to calculate false discovery rates which are dependent on *i*. The cut-off for a minimal read count per window for the peak calling is then chosen according to a user-defined maximal FDR and all windows that have a read count that exceeds this threshold are called enriched.

Finally, the applied peak calling procedure incorporates optional post-processing steps which can further enhance the peak detection. To precise peak localization, *localization boundary refinement* can be applied, meaning that reads coming from the forward and reverse strand are separated and the maxima of the individual strand-specific peaks are used to predict the true boundaries for the enriched sites instead of using the fixed window size. Moreover, *single-strand filtering* can be applied which removes peaks that are mostly based on reads from the 5' direction without corresponding 3' peaks and *vice versa*. According to Ji *et al.*, both filters will lead to better peak results when compared to no post-processing.

2.3.7 Gene Ontology (GO) Term Enrichment Analysis

In this study, GO term enrichment analysis is used to find specific GO terms that are overrepresented in a set of genes (target genes from the ChIP-chip analysis or differentially expressed genes in the siRNA knockout) when compared to all analyzed genes. The Gene Ontology[177] is an initiative that aims to standardize the representation of genes and gene annotations across species and databases. The ontology is divided into the three main compartments 'molecular function', 'cellular component' and 'biological process'. The main advantage of using GO is that is has a defined set of terms to represent individual annotations and provides a structured form for the relationship between these GO terms. This structure consists of a directed acyclic graph where each term is represented as a node and the relationship to one or more other terms is represented as directed edges. These relationships are either "is_a" or "part_of" relations, with increasing specificity of GO terms that are more distant from the

root node. Therefore, while a gene is always annotated to the most specific GO term possible, it is likewise annotated to all ancestor terms (terms that are on any root from this term to the root term).

To test overrepresentation for any GO term it would in general be possible to use the hypergeometric distribution

$$P(K=k)=\frac{\binom{g}{k}\binom{N-g}{n-k}}{\binom{N}{n}} \quad ,$$

where N is the total number of analyzed genes (background), whereof n belong to the selected set of (*e.g.* differentially expressed) genes, and g is the number of genes having the tested GO term, whereof k belong to the selected gene set. However, this approach is unaware of the inherent hierarchy of the GO graph as the significance for a single term does not incorporate the significance of any of its child terms.

Instead, the algorithm proposed by Alexa *et al.*[175] which implicitly integrates the GO graph structure was used to analyze overrepresented GO terms. The procedure is depicted in Algorithm 1. Each GO term is thereby investigated bottom up in the graph hierarchy, testing associated genes against the background set. Subsequently, genes that have already been marked as significantly associated to a GO term are removed from any further tests for all ancestor terms. Instead of using a Fisher's exact test as proposed in the original publication from Alexa *et al.* the more suitable hypergeometric test statistics was utilized as suggested by Falcon *et al*[176].

2.3.8 Prediction of Transcription Factor Binding Sites (TFBS)

In the analysis of transcription networks, the determination of true transcriptional regulators and their targets is most essential. One approach taken in this study to predict transcriptional regulators and thereby infer gene regulatory networks is the prediction of binding sites for transcription factors from sequence data. Based on the biochemical process of transcription factor binding to *cis*-regulatory elements in the promoter of their target genes descriptors of the binding behavior for a large number of TFs have been gathered.[178,179] The most common form to represent TF motifs are position weight matrices (PWMs), which represent motifs in a matrix form with one row per symbol a of the alphabet $\mathrm{A}=\{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$ and one column $i \in \{1,2,\ldots,L\}$ for each position in a pattern of length L. Each combination of symbol and position has a score assigned which typically represents the count or frequency $f_{a,i}$. As a PWM assumes independence between positions in the pattern, the score between the PWM and DNA sequence site S of the same length can be calculated as the sum of the individual symbol-position combinations. A common graphical representation for a PWM is the '*sequence logo'* supposed by Schneider and Stephens.[180] An example for a PWM, its sequence logo and real DNA binding sites is given in Figure 14. PWMs can be used to predict the binding of a TF to a promoter sequence. Two different methods have been utilized: the MATCH program uses predefined score cutoffs to predict individual binding sites for the TF under study while the TRAP approach derives affinity-based predictions incorporating whole promoters. Therefore, the two approaches differ

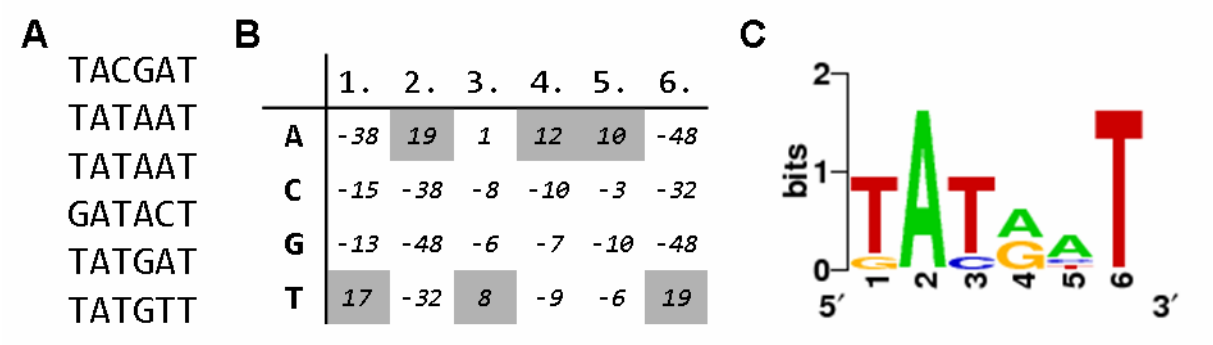

Figure 14: Different representations of a *cis*-regulatory element
(A) An example of six sequences corresponding to the -10 region of *E. coli* promoters. **(B)** PWM representation of the same region using a large number of sequences. The best scoring nucleotide is colored in gray. **(C)** Sequence logo. The example is taken from Bulyk.[182]

greatly. The advantage of TRAP against MATCH is that also contributions from weaker binding sites will be integrated, which might fall below the MATCH thresholds. However, the application of MATCH is much more straight-forward, if individual binding sites in a given sequence should be determined. The TRAP approach, on the other hand, requires multiple sequences to be present as the affinity of a single sequence without the knowledge of an overall distribution of sequence affinities is meaningless.

MATCH

The MATCH program from Kel *et al.*[181] is provided with TRANSFAC.[178] For a given PWM, MATCH computes the position specific information vector

$$I(i) = \sum_{a \in \mathrm{A}} f_{a,i} \ln(4 f_{a,i}) \quad ,$$

which is closely related to the entropy definition as introduced in section 2.3.3. A normalized and information weighted similarity score between the PWM and S is then computed using the formula

$$\textit{Similarity Score}(S, \mathrm{PWM}) = \frac{\sum_{i=1}^{L} I(i) f_{a,i} - \sum_{i=1}^{L} I(i) \min_{a \in \mathrm{A}}(f_{a,i})}{\sum_{i=1}^{L} I(i) \max_{a \in \mathrm{A}}(f_{a,i}) - \sum_{i=1}^{L} I(i) \min_{a \in \mathrm{A}}(f_{a,i})} \quad .$$

Due to the affine linear transformation the score ranges between zero and one. The advantage of incorporating the information vector is the more stringent penalization of mismatches in highly informative regions and less stringent in uninformative regions. MATCH calculates this similarity score for the whole matrix (*matrix similarity score*) as well as the five most informative positions as defined from the information vector (*core similarity score*). Correspondingly, two different thresholds need to be defined and exceeded to predict a TFBS. TRANSFAC provides matrix and core similarity thresholds for all of their matrices optimized on a set of biological sequences to reduce the number of false positives, false negatives or to minimize the sum of both which makes the MATCH program an

appropriate tool for TFBS prediction.[181] To search both the forward and the reverse strand MATCH scans both strands separately and returns all found matches.

TRAP

A complementary approach for the prediction of transcriptional regulators was integrated in the analysis of cardiac regulatory networks in human, namely, the transcription factor affinity prediction (TRAP) by Roider *et al.*[183] In the same line as the TFBS prediction by MATCH, their model is based on the definition of position specific mismatch energies between the TF (which is again represented by its PWM) and the sequence site *S*. However, Roider *et al.*[183] directly incorporate data about background frequencies of nucleotides. They model the mismatch energy $\varepsilon_{a,i}$ as

$$\varepsilon_{a,i} = \frac{1}{\psi} \log\left(\frac{\max_{a \in A}(f_{a,i})}{f_{a,i}} \cdot \frac{\max_{a \in A}(b_a)}{b_a} \right) \quad ,$$

where *b* contains the background frequencies obtained for the individual nucleotides and ψ is introduced to scale the mismatch energies in units of thermal energy.[183] The mismatch energies of a whole DNA site ε_S is again calculated as the sum over the individual mismatch energies $\varepsilon_{a,i}$ given the nucleotide sequence. Based on the model for the fraction of bound sites proposed by Zumdahl,[184] Roider *et al.* model the affinity $F(S)$ of the TF to *S* as

$$F(S) = \frac{R_0 e^{-\varepsilon_s}}{1 - R_0 e^{-\varepsilon_s}} \quad ,$$

where they define R_0 as the activity of the TF times the site-specific equilibrium constant of the site with the highest affinity. Using ChIP experiments, Roider *et al.* defined optimized values of R_0 and ψ. However, instead of calculating thresholds which separate true binding sites from unbound sites, the individual site specific affinities are summed over whole promoters to gain an overall affinity score which can be used to rank promoters in terms of their likeliness to bind a certain TF.

A main drawback of both approaches and TFBS prediction in general is the low signal-to-noise ratio which is commonly present in promoters of genes and which leads to many false-positive predictions. This drives the construction of regulatory networks solely based on TFBS prediction unusable. This problem is further aggravated by large distances between binding sites and the TSS observed in many studies which requires large promoter regions to be scanned. A common way to increase the signal-to-noise ratio is the use of conservational information. The main idea is that regions with a strong regulatory impact are positively selected against mutations. Thereby, regions that show high variability between the organism of interest and closely or distantly related species can be excluded from the prediction of functional binding sites. Therefore, the use of conservation information was integrated into the prediction of TFBS throughout this study. The two ways used to assess the sequence conservation are alignments between the sequence of interest and an orthologous sequence from a related species as well as the PhastCons score. The latter is based on a phylogenetic hidden Markov model which is trained using multiple-species alignments. PhastCons scores for genomic

regions were downloaded from the UCSC genome browser.[185] However, one has to keep in mind that there are many regulatory differences even between closely related species and every approach that highly relates on the presence of conservation will not find regulatory sites driving these differences.[186] Another way to increase the power of TFBS predictions, which was implemented in the prediction of regulatory subnetworks based on patient data, is to use a set of promoters of genes which are co-expressed and therefore likely co-regulated by common TFs. A TFBS which is predicted in all or a high number of these promoters is then a very likely candidate for a common regulator. A third approach not taken in this study is to search clusters of elements, which are often referred to as *cis*-regulatory modules. This approach is based on the finding that TFs often bind in co-occurrence with a specific set of other TFs and was successfully used in a number of studies.[187-190]

2.3.9 De Novo Motif Prediction

The prediction of *cis*-regulatory elements or modules does not *per se* depend on previously known motifs. A number of algorithms have been developed that try to predict these elements *a priori* from sequence data. As earlier publications have suggested that ChIP results can be used to enhance existing PWMs, these were further integrated into the analysis of ChIP-chip peaks (section 3.1.5). Their implementations rely on different algorithmic techniques like Gibbs sampling or heuristic based enumerations of all frequent patterns. As all these tools output a number of likely regulatory elements with unknown significance, it has previously been suggested that several of these tools should be used in any motif search and that their results should be combined and compared to already known motifs.[191,192]

To predict TFBSs *de novo*, three prediction tools were used, namely BioProspector[193], AlignACE[194] and Wedder.[195] The former two implement a modification of the Gibbs sampler for motif discovery originally proposed by Lawrence *et al.*[196] Based on an input set of sequences which should contain a common motif and the expected length of this motif, the Gibbs sampler starts with a random set of subsequence positions. In every step, the algorithm samples a new subsequence position for a single input sequence using scores based on the frequency matrix derived from all remaining input sequences while incorporating a background frequency. The common motif is found when the algorithm converges to stable subsequence positions, making it very similar to the Monte Carlo method. The two aforementioned algorithms mainly improve the originally proposed implementation by lessening the requirement of a single occurrence of the motif in every input sequence and incorporating the reverse strand. They mainly differ in the used background model and their strategy to find multiple motifs. While AlignACE uses overall GC content as background and iteratively masks subsequences belonging to previously found motifs from the next run, BioProspector incorporates a 3rd-order Markov model and starts in each run from different points in the initial search space. As an additional tool, Weeder was used as it implements a search strategy that is complementary to the two others. Weeder first builds a suffix tree[197] from all input sequences and then searches for all pattern of a given lengths with less than a given number of errors. The search space is reduced by imposing a restriction on the number of errors allowed in prefixes of the final patterns. A further tool used was MEME,[198] which performed less convincing such that only few specific motifs were received meanwhile requesting a much longer running time than the other tools. Therefore it was removed from the analyses.

As the different algorithms in general provided similar resulting motifs, these motifs were subsequently clustered hierarchically using the 'Tree' algorithm provided with the AlignACE package. As a measure of distance, the pairwise Pearson correlation coefficient for the six most informative positions (explained above for the MATCH tool) was computed between each found motif using the 'CompareACE' tool.[194] The resulting clustering was then used to define groups of very similar motifs by applying a correlation coefficient cut-off of 0.6 as proposed in the original publication. To find the most specific motif in every group each motif was scored using the *group specificity score* introduced for motif finding by Hughes *et al.*[194] This scores is based on the hypergeometric distribution and is used to calculate the probability of finding the observed or a better overlap between the sequences used to predict the motif and the total number of sequences from the input and background set that contain the motif. The group specificity score has already been used in a large number of studies[192,199,200] and was shown to successively discriminate real binding sites from background noise.[201] To finally predict a set of *de novo* motifs, only the best-scoring motif of each cluster was kept and compared to all known TF motifs from TRANSFAC again using the Pearson correlation coefficient of the six most informative positions.

2.3.10 Relational Databases

As a final step to make the results of this study available for other researchers, CARIN, the CArdiovascular Regulatory INteraction database was designed and implemented in form of a relational database. Relational databases are the most common form of databases today. They are based on a relational model for data management, which describes data in forms of relations (tables) which have a fixed number of attributes (columns) that are often given a fixed data type (*e.g.* a string). A single entry, called a tuple, is a set of attributes and a single relation consists of many tuples (rows) with the same attributes. To link data between several tables of a relational database keys are used, which consists of single attributes that are equal in tuples from different tables. Further, each table consists of a primary key which uniquely defines a certain tuple. This primary key can consists of a single attribute or a set of attributes, which is then called a surrogate. A relational scheme which depicts the implemented tables and links is used to visually represent the structure of a relational database.

To manage relational databases and retrieve data, the *Structured Query Language* (*SQL*) has been established. SQL consists of a fixed number of words to create and remove tables and tuples, pose restrictions on attributes and tables, retrieve data from one or multiple tables using keys and so on. Software used to facilitate the described functionality including SQL is often called a relational database management system (RDBMS) and many such systems have been developed. The CARIN database was implemented as an SQLite database.[202] SQLite is an embedded RDBMS that stores the whole database in one file. The main advantage of using SQLite is that it doesn't require any database installation and has supported database interface (DBI) access provided by all common programming languages including the used Perl and R.

3. Results

While a panel of important cardiac regulators has been identified which maintain correct cardiac development and function, only little is known about their interaction, their interplay with epigenetic and environmental factors or the breakdown of regulatory networks in cardiac disease. This study provides a systems-biology approach integrating a number of high-throughput genome-wide cell culture experiments with data from mouse heart time series as well as patient with congenital heart disease to study the cardiac regulatory network at a systems-level.

In an initial step, the steady-state regulatory network of four key transcription factors was analyzed using several genome-wide high-throughput datasets (section 3.1). First, it was focused on the direct downstream targets by evaluating *in-vivo* DNA-binding sites of the respective factors using ChIP-chip. Then, functional consequences of the proposed regulation in knockdown experiments were investigated and respective transcription networks were built. To determine the impact of epigenetic regulatory factors, the co-occurrence of activating histone modifications with TFBS was analyzed and related to the gene expression levels of direct targets. Key findings for the two regulatory factors Srf and H3ac were validated using ChIP-seq. The results obtained were subsequently analyzed in a time series of in mouse hearts around cardiac maturation (section 3.2). Finally, the relevance of the results gathered in cell culture and mouse hearts was analyzed using data from patients with a broad range of congenital heart disease and combined with the detection of disease-associated profiles. To prepare the ground for future studies, CARIN, the CArdiac Regulatory INteraction database was designed and implemented (section 3.4).

Results obtained in the analysis of the individual datasets will be integrated into the discussion (section 4).

3.1 Combinatorial Regulation of Four Transcription Factors and Accompanying Histone Modifications

3.1.1 ChIP-chip Data Normalization and Peak Calling

The *in-vivo* binding site data consisted of ChIP enrichment intensities of the four TFs Gata4, Mef2a, Nkx2.5 and Srf which were measured by microarray detection. It comprised two independent ChIP pools as well as a non-enriched Input sample for each analyzed TF. See section 2.2.1 for a full description of the ChIP-chip dataset. The initial steps in the analysis of the ChIP-chip experiments were data normalization to remove systematic biases and the appliance of ChIP-chip peak calling.

Normalization of the ChIP-chip Data

To normalize the ChIP-chip datasets, the variance stabilization normalization[113] (vsn, section 2.3.2) was used as the inspection of MA and MA_{rank}-plots of pairwise unnormalized ChIP-chip data revealed a clear variance dependency of the signal intensity (Figure 15). Vsn provides a parametric solution for this dependency and simultaneously normalizes systematic biases between experiments. Applying vsn significantly reduced the variance dependency and also the 'banana-shape' of the MA- plots (Figure 16). As vsn doesn't integrate a probe sequence specific normalization, it was checked whether

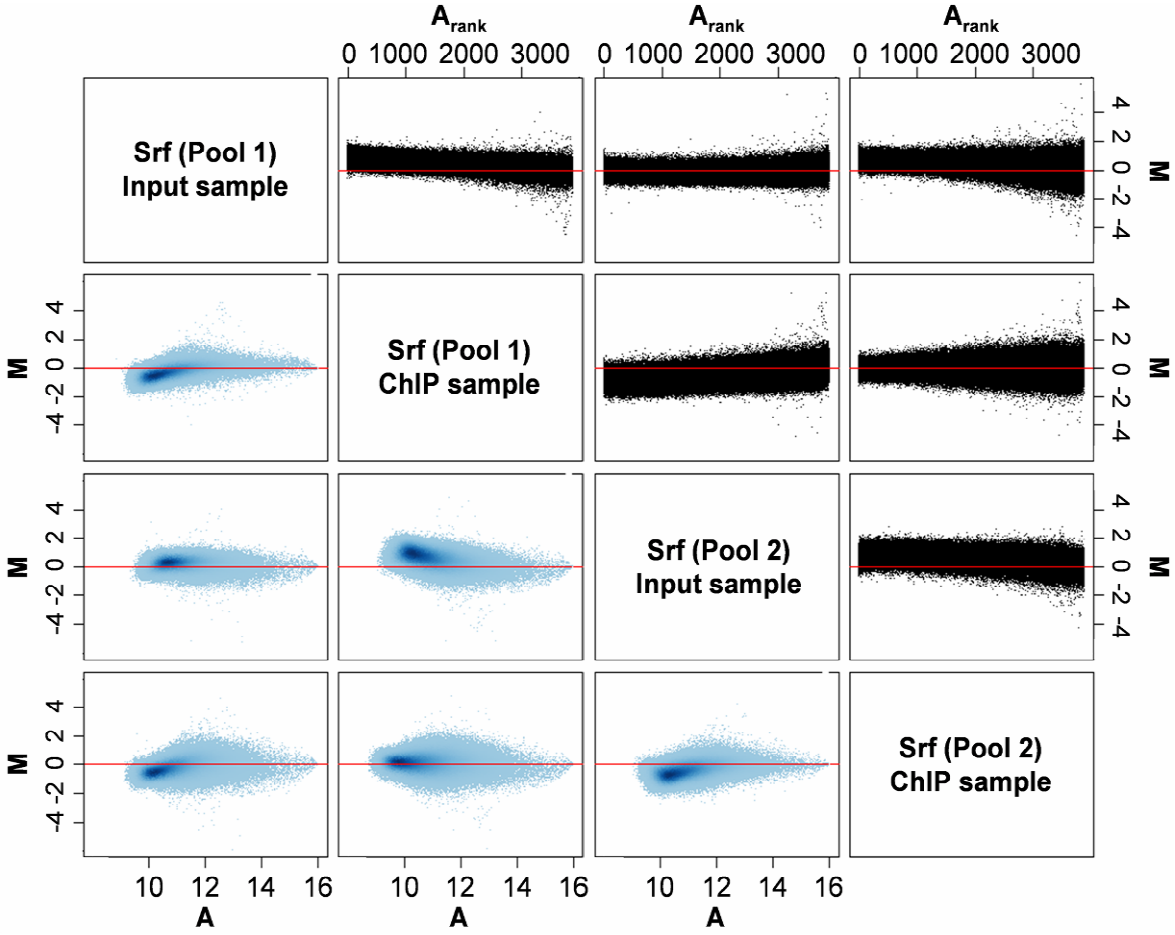

Figure 15: MA (blue) and MA_{rank} (black) plots for SRF ChIP-chip data before normalization

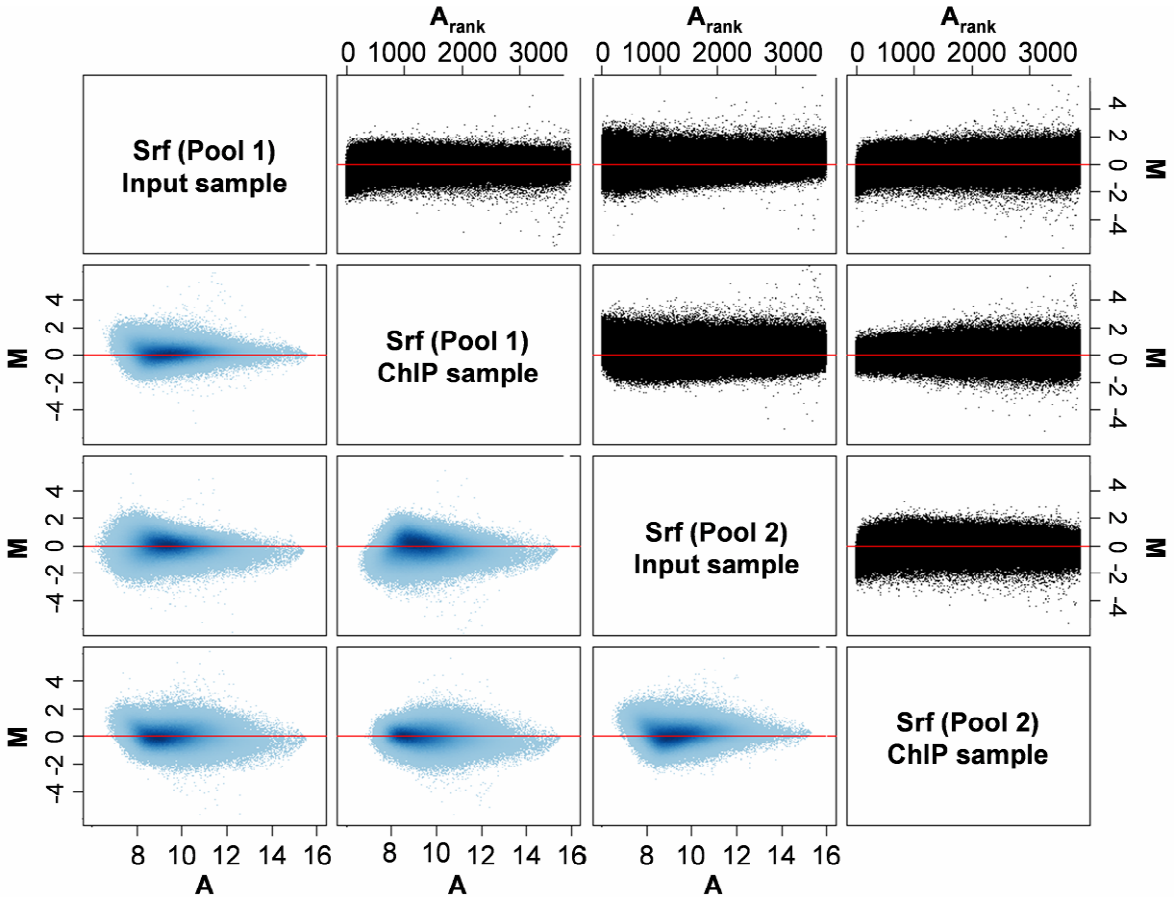

Figure 16: MA (blue) and MA_{rank} (black) plots for SRF ChIP-chip data after vsn normalization

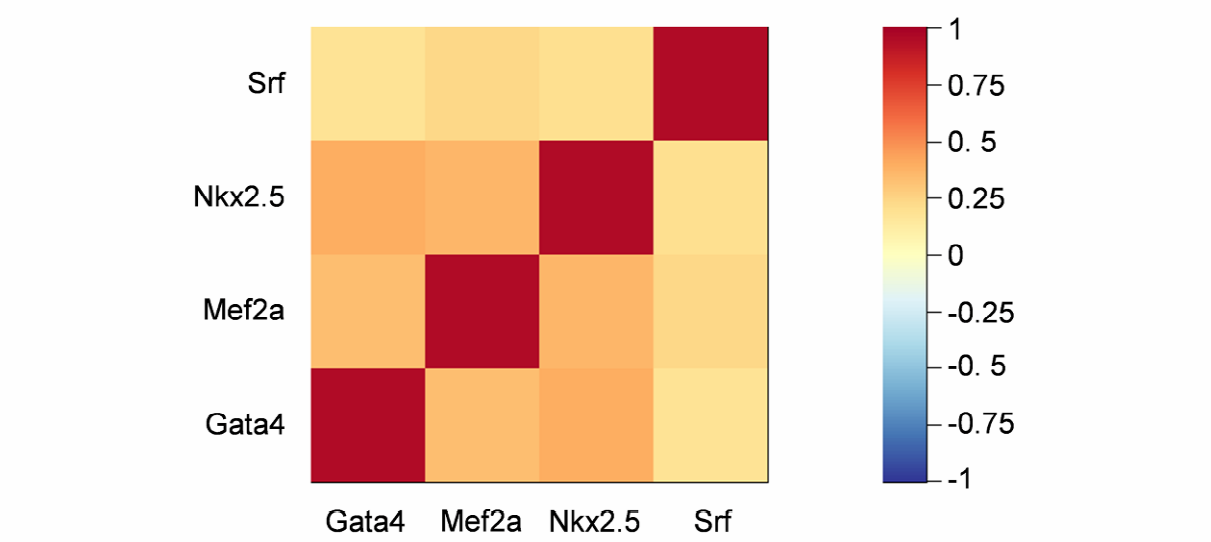

Figure 17: Correlation between ChIP-chip probe levels of individual TFs
The heatmap indicates pairwise Pearson correlation coefficients between the enrichment levels of all measured probes. The legend on the right depicts the color coding.

correlations between the probe intensities of the four analyzed factors exist, which could point to a sequence dependent bias. Computing pairwise Pearson correlation coefficients over probe intensities no prominent correlation between individual probe intensities for any two TFs was found(Figure 17). Further, the correlation between the GC content of probes and their mean probe level was determined. Again, only a negligible weak correlation was found (Pearson correlation coefficient of 0.184). Finally, log-ratio enrichment levels for each probe were computed by subtracting log Cy3 (Input) from log Cy 5 (ChIP sample) for every probe and every ChIP pool, resulting in two enrichment values per probe for each TF (one for each pool).

ChIP-chip peak calling

Using a peak calling algorithm based on an empirically derived distribution of unbound probe intensities (section 2.3.5) followed by multiple testing correction using the *Benjamini–Yekutieli*[148] procedure (section 2.3.4) to control the false discovery rate, several hundred of binding sites were identified for each TF. Thereof, Srf had the most peaks (1,335) followed by Mef2a (999), while the binding pattern of Gata4 and Nkx2.5 was more specific with 447 and 383 peaks each, respectively. Subsequently, the called ChIP-chip TF peaks were assigned to genes if they lay less than 10 kb upstream or inside a gene (gene annotation taken from Ensembl[106] version 45) in accordance with the definition of the genomic positions represented on the ChIP array. In line with their high number of peaks, Srf (1,150) and Mef2a (701) had the most target genes, while lesser targets were found for Gata4 (345) and Nkx2.5 (276) The number of peaks and associated target genes is further depicted in Table 3.

	Gata4	**Mef2a**	**Nkx2.5**	**Srf**
No. of peaks	447	999	383	1,335
Target genes	345	701	276	1,150

Table 3: Number of ChIP-chip peaks and related target genes for every TF

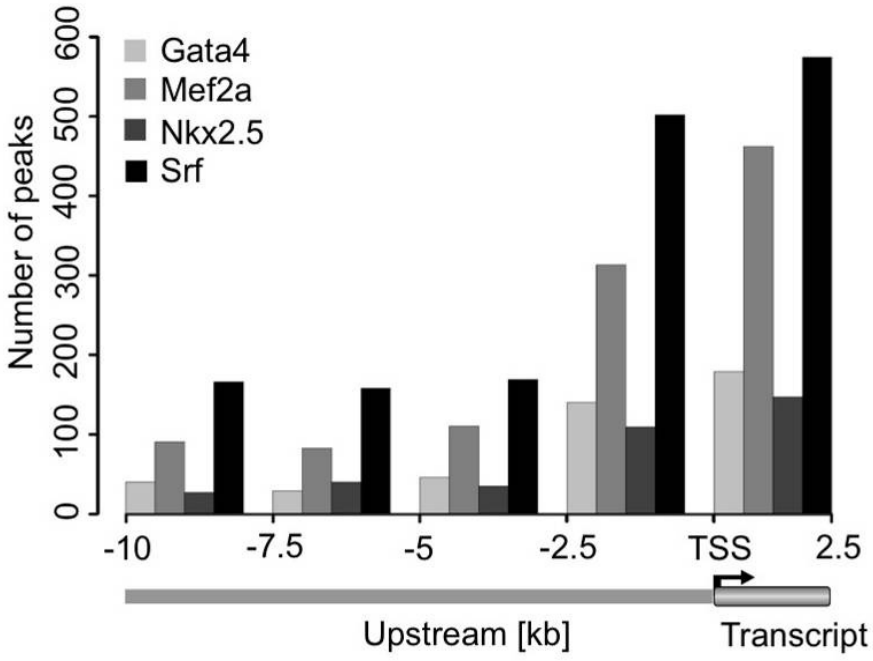

Figure 18: Positional distribution ChIP peaks relative to the TSS
The y-axis shows the number of transcription factor binding sites per transcription factor as bar plots in 2.5 kb windows.

3.1.2 Positional Distribution of Found ChIP-chip Peaks

As an initial analysis the genomic binding pattern of the four TF measured in ChIP-chip were investigated. Looking at the histogram of distances to the next TSS these binding patterns were found to be very similar. Most of the binding sites (76%) reside very close to a TSS in a window of ± 2.5 kb (Figure 18). While the found distribution could partially reflect potential biases by assigning only the closest TSS or to a lesser extent the design of the ChIP array, the same behavior was found in other studies analyzing TF binding behavior, leading to the definition of a core promoter binding region.[203,204] However, another 24% of TFBS were found between 2.5 kb and 10 kb distant to any annotated TSS, showing a uniform localization in these potential enhancer regions.

3.1.3 Gene Ontology Analysis

To evaluate the reliability of the used ChIP-chip peak calling and to gain insight into the functionality of each of the individual factors a Gene Ontology term enrichment analysis was conducted using the biological process subtree of terms and applying the algorithmic approach by Alexa *et al.*[175] as described in section 2.3.6. All genes represented on the ChIP array were taken as reference.

The resulting significant GO terms reflect the importance of all four TFs for heart and muscle development, even when compared to the heart/muscle specific background as given by the array design. Table 4 to Table 7 show only the most significant GO terms (p-value < 5x10-5) for each ChIP experiment. Furthermore, the resulting significant GO terms were highly related to the phenotypes reported for the respective transcription factor. For example, the GO terms '*muscle contraction*' and '*heart looping*' are significantly overrepresented among Mef2a and Nkx2.5 targets, respectively, and

GO Term	GO ID	p-value
heart development	GO:0007507	$2.1x10^{-6}$
striated muscle development	GO:0007519	$1.9x10^{-5}$
skeletal muscle fiber development	GO:0048741	$2.6x10^{-5}$
positive regulation of cell proliferation	GO:0008284	$2.9x10^{-5}$
muscle contraction	GO:0006936	$4.4x10^{-5}$

Table 4: GO term analysis of genes bound by Gata4 ($p<5x10^{-5}$)

GO Term	GO ID	p-value
muscle contraction	GO:0006936	$1.0x10^{-8}$
actin cytoskeleton organization and biogenesis	GO:0030036	$1.2x10^{-6}$
cytoskeleton organization and biogenesis	GO:0007010	$2.4x10^{-6}$
adult heart development	GO:0007512	$5.6x10^{-6}$
heart development	GO:0007507	$6.8x10^{-6}$
circulation	GO:0008015	$7.9x10^{-6}$
cardiac muscle development	GO:0048738	$4.5x10^{-5}$

Table 5: GO term analysis of genes bound by Mef2a ($p<5x10^{-5}$)

GO Term	GO ID	p-value
heart development	GO:0007507	$1.6x10^{-6}$
cardiac inotropy	GO:0002026	$6.5x10^{-5}$
cell adhesion	GO:0007155	$1.2x10^{-4}$
neural crest cell development	GO:0014032	$2.6x10^{-4}$
muscle contraction	GO:0006936	$2.8x10^{-4}$
negative regulation of heart contraction	GO:0045822	$3.1x10^{-4}$
bone mineralization	GO:0030282	$4.1x10^{-4}$
heart looping	GO:0001947	$4.1x10^{-4}$
cell motility	GO:0006928	$4.3x10^{-4}$

Table 6: GO term analysis of genes bound by Nkx2.5 ($p<5x10^{-5}$)

GO Term	GO ID	p-value
biological regulation	GO:0065007	$1.7x10^{-6}$
regulation of cellular process	GO:0050794	$2.5x10^{-6}$
transcription	GO:0006350	$4.9x10^{-6}$
cellular developmental process	GO:0048869	$9.1x10^{-6}$
muscle contraction	GO:0006936	$1.6x10^{-5}$
regulation of heart contraction	GO:0008016	$1.6x10^{-5}$
regulation of metabolic process	GO:0019222	$2.0x10^{-5}$
regulation of transcription, DNA-dependent	GO:0006355	$3.6x10^{-5}$
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0019219	$4.0x10^{-5}$
RNA biosynthetic process	GO:0032774	$4.4x10^{-5}$

Table 7: GO term analysis of genes bound by Srf ($p<5x10^{-5}$)

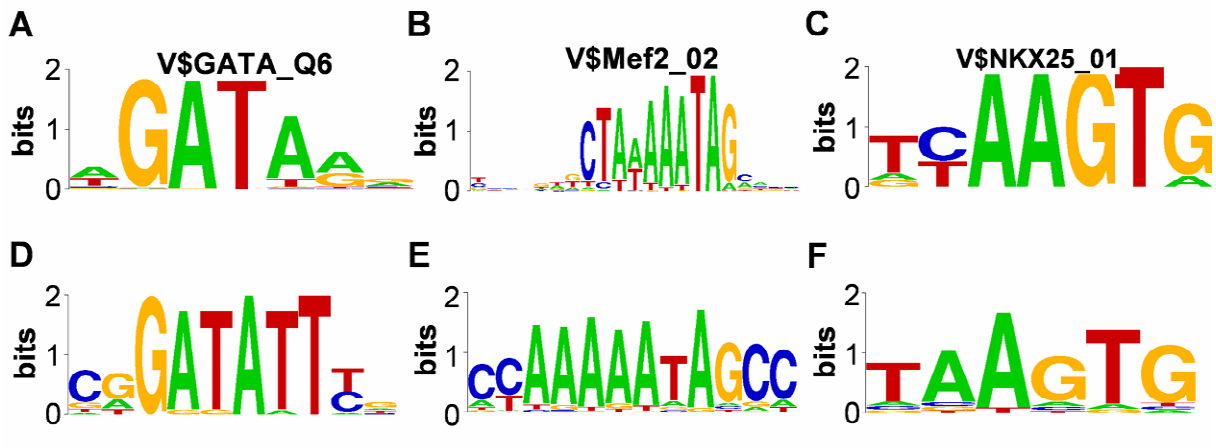

Figure 19: Previously described and *de novo* motifs from ChIP-chip peaks
(A-C) Previously described TF matrices taken from TRANSFAC for **(A)** Gata4, **(B)** Mef2a and **(C)** Nkx2.5. **(D-F)** *De novo* found motifs for **(D)** Gata4, **(E)** Mef2a and **(F)** Nkx2.5. No *de novo* motif could be identified that resembled a TRANSFAC matrix associated to Srf.

both are key features of corresponding knockout mouse models.[52,205] This findings support the validity of the implemented ChIP-chip peak calling approach.

3.1.4 *De Novo* Motif Prediction

To proof the validity of the found peaks and to check whether the ChIP-chip dataset could be used to improve existing PWMs, a *de novo* motif discovery approach was conducted as described in section 2.3.9. To reduce the amount of sequence potentially unrelated to the TF binding, only the sequence ±250 bp from each peak's center was used as input for the prediction. In addition to the found ChIP-chip peaks, a *negative* peak set was defined that contained 100 regions represented on the ChIP-array that had the smallest standard deviation and a mean of approximately zero. For those algorithms that provided the possibility to use a background set to estimate background nucleotide frequencies, this negative peak set was used. For each TF the best 10 assembled motifs were collected. Using this approach the known motifs for Gata4 and Nkx2.5 could be re-identified as predominant motifs in their respective peaks but not for Mef2a and Srf.

In an attempt to further increase the signal-to-noise ratio in the used peak sequences, masking of conserved regions was performed using the full BlastZ alignment between mouse and human that was retrieved from Ensembl (mouse assembly NCBI m37, human assembly NCBI 36). In addition to the single nucleotide conservation masking provided by the alignment, a 100 bp window was shifted along the peaks and windows exceeding 70% of conservation remained unmasked. Using the masked peak sequences resulted in the finding of the Mef2a binding motif, but was still not able to detect the Srf motif known from TRANSFAC[103] (Figure 19).

3.1.5 Binding Site Prediction Using Known Motifs

After the *de novo* identification of binding sites was only partially successful, the sequence underlying the transcription factor binding sites were analyzed in more detail now directly searching for TRANSFAC motifs within the presumably bound sequences ±250 bp of the center of each peak. All

	Gata4	Mef2a	Nkx2.5	Srf
Total number of peaks	447	999	383	1,335
Peaks containing predicted TFBS	421	858	323	169
Predicted TFBS conserved to human	139	148	111	65
Predicted TFBS conserved according to PhastCons	122	267	103	51

Table 8: Number of ChIP-chip peaks with predicted TFBS and conservation information

matrices from TRANSFAC that were associated to the four analyzed factors were searched using the TRANSFAC MATCH[181] algorithm (section 2.3.8) together with the predefined cut-offs for core and matrix similarity to reduce both the type I and type II error levels. The results indicated a high percentage of peaks (84-94%, Table 8) with appropriate binding sites for Gata4, Nkx2.5 and Mef2a.

To identify to which extent the TFBSs are conserved it was investigated how often the predicted binding sites occurred in conserved regions. Therefore, two measures of conservation were used, first, single base-pair conservation between human and mouse using the full BlastZ alignment between mouse and human as described before, and second, conservation between 18 vertebrate species as based on PhastCons elements (section 2.3.8) that were retrieved from the UCSC Genome browser.[185] Only 15-31% of Gata4, Nkx2.5 and Mef2a peaks had predicted binding sites which were found to be completely conserved between mouse and human and only ~27% lay in conserved regions according to PhastCons elements. Thus, by focusing on conserved sequence regions alone, *a priori* more than two-third of potential TFBS would be missed.

For Srf, the fraction of ChIP-chip binding events that harbored predicted TFBS was very small with 169 out of 1,335 peaks. In addition, the Srf motif could not be recovered in the *de* novo approach. As both methods rely on Srf PWMs, the lower number of binding sites found with the *de novo* and TRANSFAC motif search approaches might reflect a potentially insufficient representation of the real Srf binding. However, Srf is well-known to bind the CArG-box motif $CC(A/T)_6GG$.[53] Therefore in a last attempt to search for Srf binding sites, a pattern matching approach was conducted to search the exact CArG-box pattern in Srf ChIP-chip peaks. Using a very relaxed setup allowing two errors at most, the CArG-box motif could be located in 1,063 (~80%) of all peak sequences, thereby providing an explanation for the Srf binding observed in ChIP. A further explanation would be Srf-binding to co-regulators which relaxes the need for the presence of Srf binding sites. Inspecting the result of the *de novo* prediction for binding sites of possible co-regulators resulted in no satisfactory candidates.

3.1.6 Combinatorial Regulation by Multiple Transcription Factors

The investigated transcription factors are known to co-regulate targets and pairwise physical interactions have been described for several of them. Nevertheless, it is unknown how frequently this co-binding occurs *in vivo*. Consequently, the assignment of Gata4, Mef2a, Nkx2.5 and Srf to the same gene was investigated. Co-binding was observed in the promoters of 498 genes, whereof 91 target genes were bound by all four transcription factors, 121 target genes were bound by three and 286 target genes were bound by two transcription factors (Figure 20 A). Looking at pair-wise co-binding odds ratios (section 2.3.3) were computed from pairwise contingency tables of bound and unbound genes (Figure 20 B). Gata4 and Nkx2.5 had the lowest number of targets (345 for Gata4 and 276 for Nkx2.5) but were observed to co-bind to 143 genes and were therefore highly correlated. In contrast,

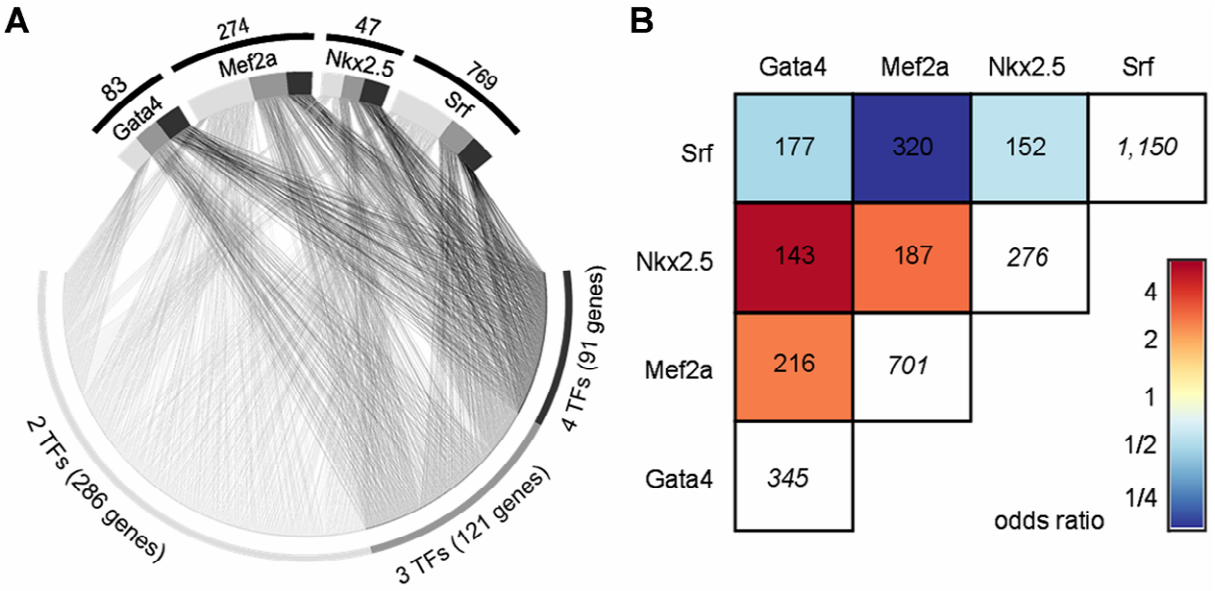

Figure 20: Co-occurrence of Gata4, Mef2a, Nkx2.5 and Srf
(A) Shown is the combinatorial binding of all four transcription factors to 498 target genes. 91 targets were bound by all four factors (black), 121 targets were bound by three (dark gray) and 286 targets were bound by two transcription factors (gray). Total numbers of genes bound by only a single TF factor are indicated above the TF. **(B)** Odds ratios of pair-wise contingency table of the occurrences of transcription factor binding sites at one gene. Total numbers of pair-wise occurrences are given. The numbers in white boxes represent the total number of bound genes for the respective TF. The odds ratio is color coded with red indicating positive and blue negative correlation.

although Mef2a and Srf show co-binding at 320 genes, they each have a much higher number of targets, leading to a very low odds ratio.

As all factors but Mef2a and Srf have been described as physically interacting, it was investigated how often two or more ChIP-chip enriched loci are observed within a distance of maximally 500 bp (Table 9). While this situation frequently occurred for two different TFs (*e.g.* 226 times for Mef2a and Nkx2.5), multiple binding sites for the same TF within 500 bp were comparatively rare (*e.g.* 22 times for Gata4). However, it is likely that many instances of multiple binding of one TF are only detected as one enriched locus due to the limited resolution of the array and the long fragment length of the ChIP-chip experiment.

3.1.7 Expression Data Normalization

To assess the functional consequences of transcription factor binding genome-wide expression measurements of transcripts were derived using microarrays. The dataset includes transcript expression of wildtype HL-1 cell as well as siRNA knockdown of each of the four analyzed TFs and two non-specific siRNA (siNon). Experiments were performed in duplicates and the four siRNA

	Srf	**Nkx2.5**	**Mef2a**	**Gata4**
Gata4	162	163	232	22
Mef2a	291	226	21	
Nkx2.5	151	11		
Srf	50			

Table 9: Co-binding of TFs within a maximal distance of 500 bp

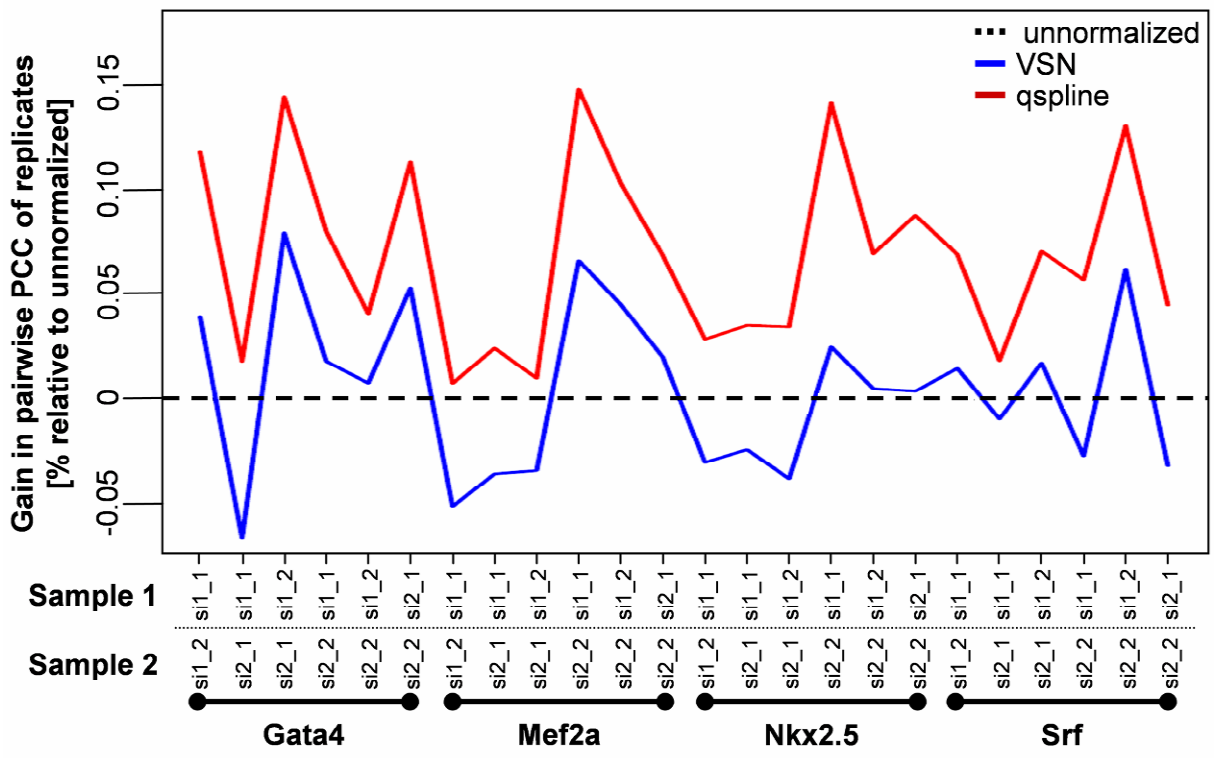

Figure 21: Comparing Pearson correlation coefficients of replicates after normalization of expression data
Shown is the gain in Pearson correlation coefficients after vsn (blue) and qspline (red) normalization compared to unnormalized expression data (black line). A negative gain indicates a decrease in the correlation coefficient. The gain is given in percent relative to the unnormalized data. SiRNA and replicate number of the individual samples are listed below the plot.

knockdown experiments were performed using two independent siRNAs for each TF, leading to a total of four siRNA expression measurements per TF. See section 2.2.2 for a full description of the dataset.

As suggested by the array manufacturer probes were initially filtered out that showed expression values lower than background based on the '*detection score*' using a number of negative control probes that are spotted on each array. This Illumina detection score $S_{Detection}$ is defined as

$$S_{Detection} = \frac{\text{rank}(\text{probe signal})}{\text{number of negative probes}} \quad ,$$

where the rank of the probe signal is computed relative to the negative control probes. According to the manufacturer's protocol, only probes with a detection score greater or equal to 0.95 in at least one experiment were retained.

Like for the ChIP-chip data, vsn was used to normalize the individual datasets. However, large inconsistencies were found between the two different siRNAs resulting in many insignificant results in the subsequent definition of differentially expressed genes. Inspecting Pearson correlation coefficients between individual replicates revealed that while the correlation coefficients of the unnormalized expression data were in general very high application of vsn frequently resulted in a reduce of the correlation between individual replicates (Figure 21 blue line). Concluding that vsn was not applicable for this expression dataset an additional non-parametric normalization method was used, namely

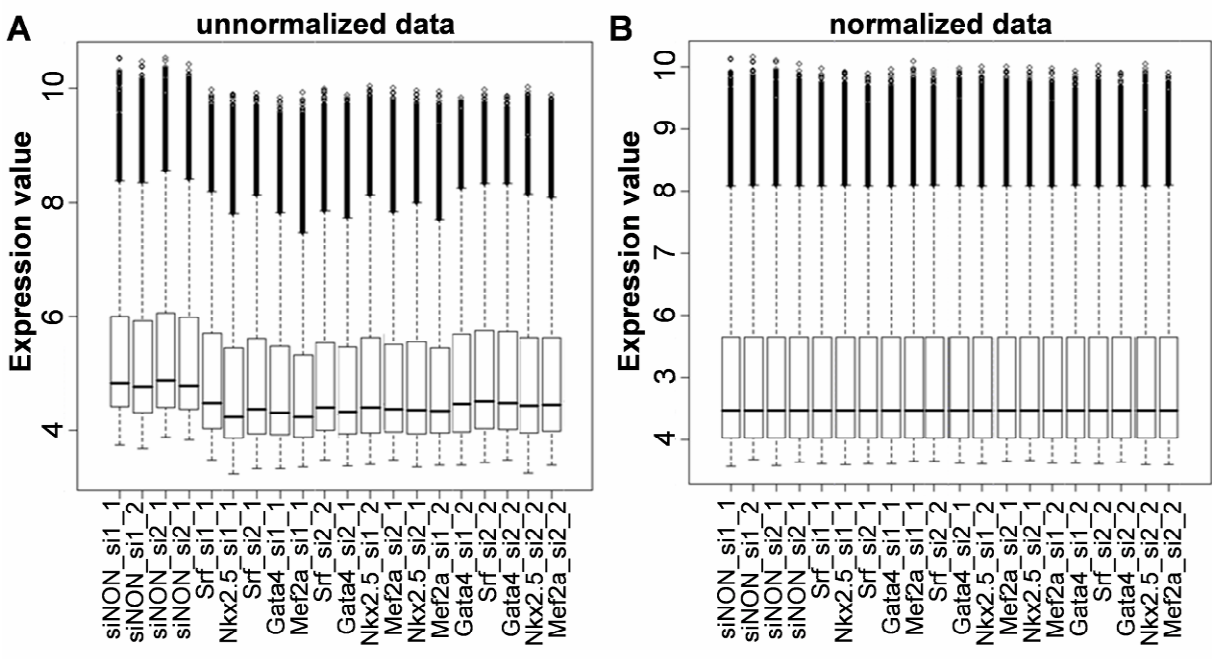

Figure 22: Normalization of expression data
Boxplots of Illumina expression values for the individual measured arrays **(A)** before and **(B)** after qspline normalization.

qspline normalization (section 2.3.2). Applying qspline to the expression data resulted in higher correlation coefficients for all replicates (Figure 21 red line) and was therefore used to normalize the expression data. Figure 22 shows the resulting boxplots for each measured experiment after qspline normalization.

In contrast to the ChIP array, which had a probe length of 70 nucleotides and was specifically designed for this analysis, gene expression measurements were performed using a standard expression array with a much smaller probe length of 35 nucleotides. To remove the probe sequence bias frequently shown for these expression arrays,[206] the median polish algorithm proposed by Tukey[207] and introduced to gene array analysis by Irizarry *et al.*[208] was applied. Using all measured experiments, the algorithm fits for every probe of every transcript the linear model (section 2.3.1)

$$Y = \beta_{\text{probe affinity}} + \beta_{\text{expression}} + \varepsilon \quad ,$$

where Y is the measured expression and $\beta_{\text{probe affinity}}$ and $\beta_{\text{expression}}$ are the coefficients that divide this expression into a probe specific effect (the base line expression of each individual probe) and the effect of the individual experiments. This *polished* expression estimate was subsequently used to combine the measurements of all probes assigned to the same transcripts.

3.1.8 Transcriptional Consequences of TF Binding

To analyze whether the investigated transcription factors act mainly as activators or repressors, transcripts were first classified into *expressed* (transcript contains at least one probe with Illumina detection score greater or equal to 0.95) and *non-expressed* (all other) based on untreated cells which lead to around 80% of expressed targets for each factor. The median expression levels of all expressed

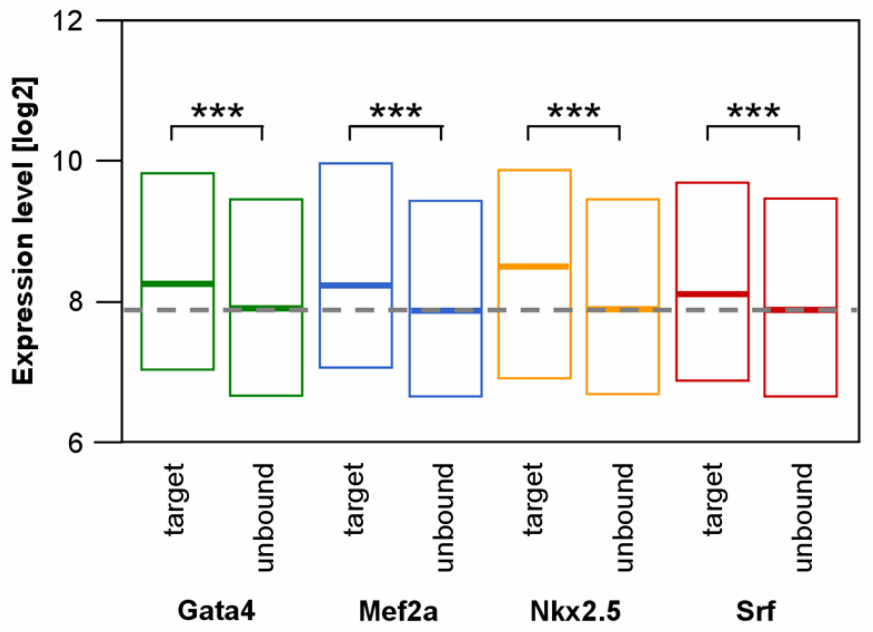

Figure 23: Boxplot of expression levels for bound and unbound transcripts.
Transcripts were grouped according to whether binding of a TF was observed in ChIP-chip (target) or not (unbound). Significant differences according to Wilcoxon rank sum p-values are indicated: ***: $p < 0.005$.

transcripts identified to be TF targets in the ChIP-chip analysis were subsequently compared to the median expression levels of all unbound transcripts represented on the ChIP and expression arrays (Figure 23). Wilcoxon rank sum tests demonstrated that the distribution of target gene expression levels is significantly elevated compared to non targets ($p < 0.005$) suggesting an overall activating function for all four investigated transcription factors.

Considering the redundant co-binding of the investigated TFs, the functional consequences of a large reduction of their individual quantity achieved by siRNA was analyzed. To define which transcripts showed a significant alteration in expression levels after the knockdown of each individual TF, the limma method developed by Smyth[209] was applied. Limma estimates for every gene a linear model which predicts the measured gene expression using the performed replicates and different expression sources (wildtype and knockdown). The advantage of estimating one linear model for each gene instead of a single model for all genes is that a single model would assume a common variance whereas multiple models can accommodate different variances. To reveal a more stable inference and higher statistical power, especially given the small number of arrays, limma's empirical Bayes method was applied to moderate the standard errors of the estimated log fold changes using the linear model fits of all genes. After the estimation, contrasts were used to compute the p-value for the comparison wildtype against siRNA knockdown.

To cope with the multiple testing problem resulting from the high number of statistical tests (one for each transcript), p-value correction was performed again applying the *Benjamini–Yekutieli* procedure. Finally, to ensure a low number of false positives, only transcripts that had a FDR-corrected p-value < 0.05 in both siRNA-mediated knockdowns when compared to siNon cells were considered to be significantly differentially expressed. Table 10 shows the resulting number of differentially up and down regulated transcripts after TF knockdown confirming a mainly activating function of all measured TFs as 73-90% of the genes were found to be down regulated in the respective knockdown. In case of Srf and Mef2a more direct targets than differentially expressed genes were found. This was

	Gata4	Mef2a	Nkx2.5	Srf
Total number of differentially expressed transcripts	621	119	782	519
Thereof down regulated	446	106	643	468
Thereof up regulated	175	13	139	51

Table 10: Number of differentially expressed transcripts in siRNA knockdown for each TF

most prominent for Mef2a, which had a high number of 701 target genes but its knockdown led to only 119 differentially expressed genes. An explanation for this finding is given by the other Mef2 family members as these are well-known to at least partially take over the function of Mef2a. As a direct proof for the performed data analysis a broad panel of differentially expressed genes could be confirmed using real-time PCR.

Analogous to the analysis of common downstream targets based on ChIP, common differentially expressed genes were determined and analyzed using pairwise odds ratios (Figure 24). While Mef2a has only a very small number of 119 differentially expressed genes in its knockdown, 92 of these transcripts are also differentially expressed in the Srf knockdown leading to a very high correlation. In contrast, while Nkx2.5 and Gata4 share 347 common differentially expressed genes, both have a much higher number of deregulated genes, leading to a very low odds ratio.

3.1.9 Combining ChIP-chip and Knockdown Results

Finally, the genes differentially expressed in the siRNA knockdown experiments and identified to be direct targets in ChIP-chip were combined. Figure 25 A shows the combinatorial regulation of a selection of heart and muscle relevant, directly bound and differentially expressed genes, which were confirmed using qPCR. This includes genes coding for structural proteins like *Actc1*, *Actn2*, *Tnnt2*, *Mybpc6*, or *Myh6*; growth factors like *Igf1* or apoptosis factors like *Casp3*. The transcription factor *Tbx20* represents an example for a gene that is bound and regulated by all four factors.

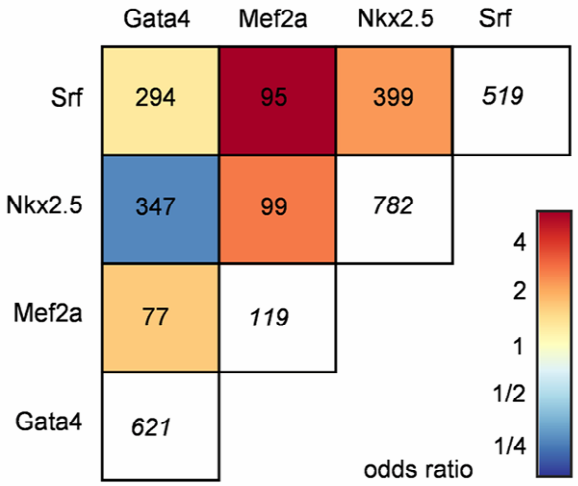

Figure 24: Functional siRNA induced knockdown of Gata4, Mef2a, Nkx2.5 and Srf
Odds ratios of pair-wise contingency table of differentially expressed transcripts after siRNA knockdown of the respective TF. Total numbers of pair-wise occurrences are given. The numbers in white boxes represent the total number of deregulated transcripts. The odds ratio is color coded with red indicating positive and blue negative correlation.

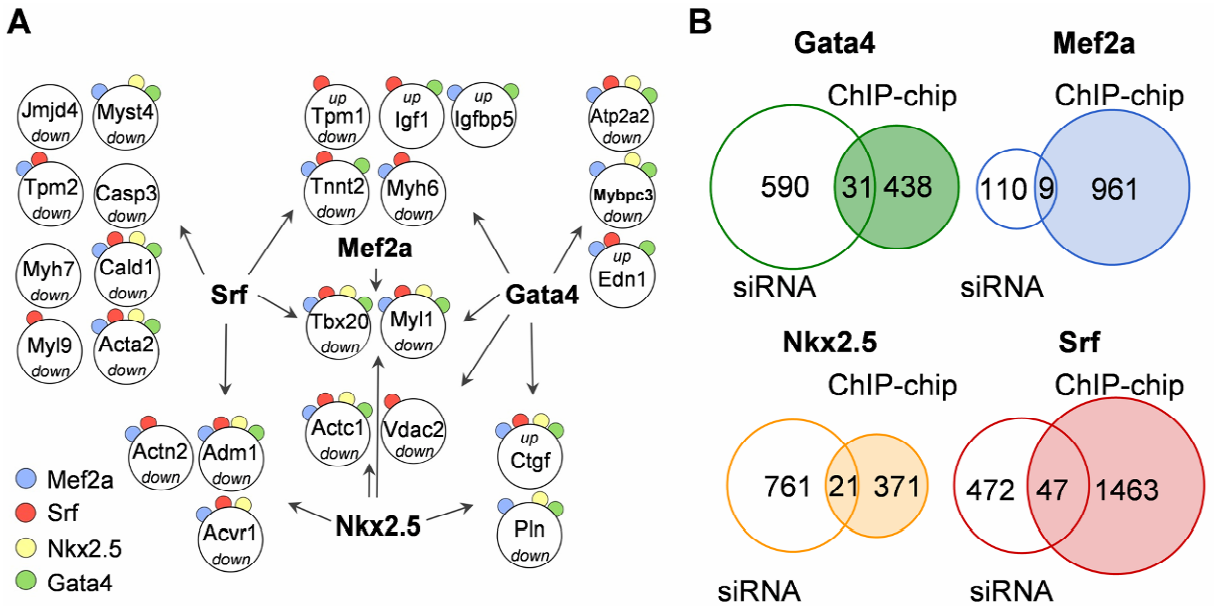

Figure 25: Combining ChIP-chip and siRNA knockdown results
(A) Transcription factor network showing a selection of cardiac relevant genes bound in ChIP-chip and significantly differentially expressed in siRNA knockdown experiment of the respective factor. Up and down regulation of genes is depicted and occurrence of ChIP binding is marked by color-coded circles. **(B)** Overlap between direct targets as measured by ChIP-chip (shaded circle) and differentially expressed transcripts as found in siRNA knockdown (blank circle).

Analyzing the overlap focusing on the functional role of the respective genes, both datasets were found to share the GO terms reflecting heart and muscle development. For example, 'muscle contraction' and 'regulation of heart contraction' are main functional roles for direct Srf as well as the respective differentially expressed genes. Interestingly, only a small fraction of direct target genes (~10%) was found to be also differentially expressed (Figure 25 B), indicating that many TFBS are occupied by TFs that may be bound in a poised state or that additional cofactors maybe lacking. Studying the literature revealed a number of different studies showing similar results such that changing the level of a factor alters the expression level of only 1-15% of the potential target genes and *vice versa* depending on the transcription factor.[210-218]

Based on the frequent co-binding observed in the study of the TF binding patterns, the possibility of a combinatorial nature of gene regulation by the four measured TFs was analyzed. Such a combinatorial regulatory influence even has the potential to explain the buffering effects found in their individual knockdowns. In accordance, genes bound by multiple transcription factors were significantly less likely differentially expressed in the siRNA experiments (χ^2-test, $p < 0.001$). Likewise, transcription factors having a high number of common binding targets share only a small number of differentially expressed genes in siRNA knockdown and *vice versa* (the correlation shown in Figure 24 is inverse to the correlation in Figure 20 B). In addition, binding in a poised state or buffering by epigenetic mechanisms such as histone modifications which infer with the accessibility of the DNA should be considered. It has to be kept in mind that transcription factor binding depends on binding affinity and accessibility of binding sites. The regulatory potential of several factors has been reported to be strongly dosage dependent (e.g. Tbx5 and Gata4). Furthermore, a significant proportion of differentially expressed genes in RNAi are likely to be regulated in an indirect manner.

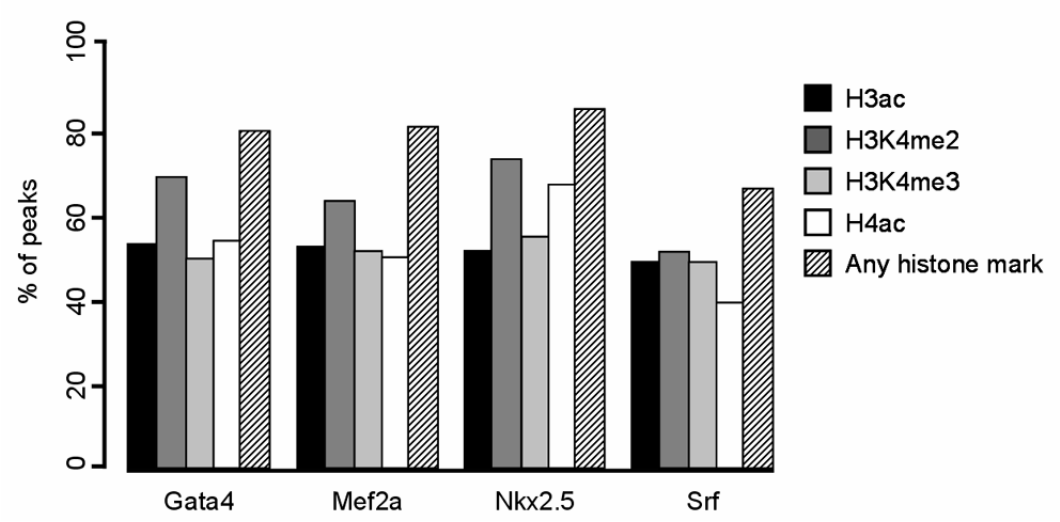

Figure 26: Overlap between histone modified sites and ChIP-chip peaks
Peaks were only considered if the respective sequences were sufficiently represented on the histone ChIP-chip array.

3.1.10 Overlap of ChIP-chip Peaks and Modified Histones

The accessibility of transcription factor binding sites is determined by the degree of DNA packaging resulting from different chromatin configurations. In turn, this is strongly influenced by histone tail modifications. Therefore, the ChIP-chip and siRNA data was analyzed in the context of co-occurring histone marks to explore the influence of histone modifications as an epigenetic mechanism to modulate cardiac gene expression. Therefore, data of the four histone modifications H3K9K14ac (H3ac), H4K5K8K12K16ac (H4ac), H3K4me2 and H3K4me3 were integrated. These histone modifications are known to promote an open chromatin state,[5,96-99] and were previously analyzed in our own lab also using ChIP-chip techniques and linear modeling.[16]

It was investigated to which extent the binding sites of Gata4, Mef2a, Nkx2.5 and Srf occurred at sites of histone modifications. As our previous study regarding the histone modifications used a smaller array set-up the number of all ChIP-chip peaks was initially reduced to those that were sufficiently represented on both arrays. Like for the combinatorial regulation by multiple TFs a histone modified sites was assigned to a TF peak if it lies within a distance of maximally 500 bp. To estimate how much overlap would be expected in a random situation, each peak was randomly repositioned onto a genomic region that was at least to a similar degree represented on both arrays as the original TF peak region while keeping the original peak length. The average of the estimated percentages of overlap after 100 repeated random associations was between 23% and 38% depending on the histone modifications (Table 11). The actual number of peaks that overlap with histone modified sites was found to be more than twice as high (65% to 84%), indicating a preferential binding at promoters marked by one or more of the investigated histone modifications (Figure 26).

	Gata4	Mef2a	Nkx2.5	Srf
Total number of TF peaks	287	592	227	734
% Overlapping	79	80	84	65
% Expected at random	30 ± 6	32 ± 3	32 ± 5	30 ± 4

Table 11: Overlap between histone modified sites and ChIP-chip peaks
The expected percentage is based on a 100-times random distribution of TF peaks on genomic sequences with marked histone modified sites.

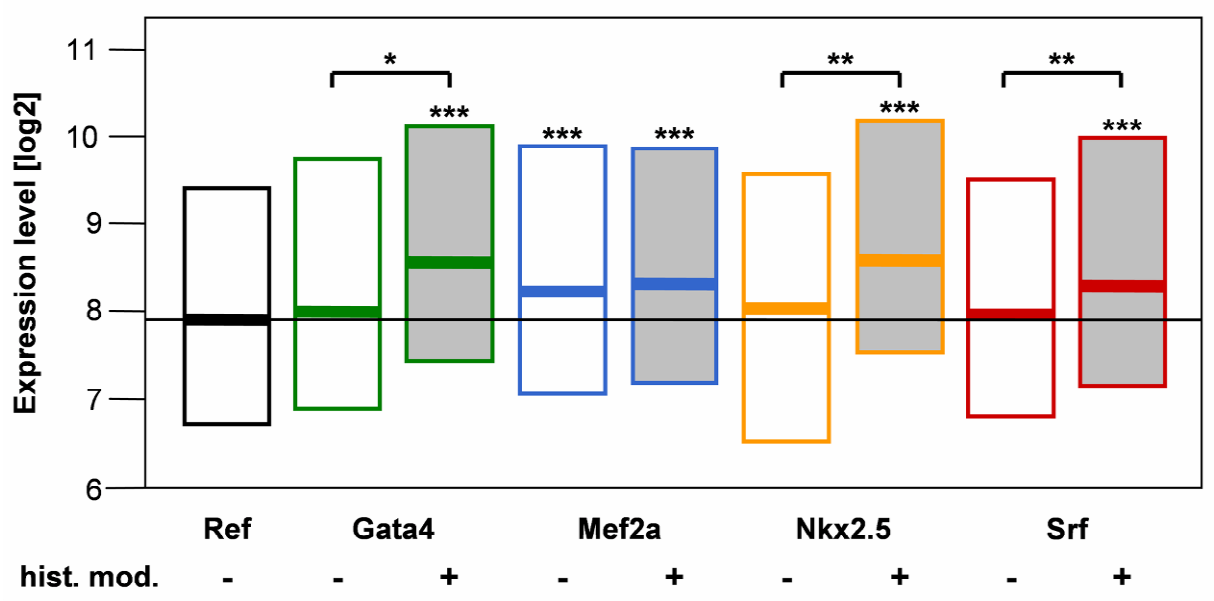

Figure 27: Influence of histone modifications on TF target gene expression
Binding sites for each TF were categorized into two groups depending on co-occurrence with a histone modification in ChIP-chip. Transcripts showing neither TF binding nor histone modified sites were used as a reference (Ref). Stars on top of each box indicate significant difference from the reference group. Stars between two boxes indicate a significant difference between the two groups. Significance levels are depicted as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$ (Wilcoxon rank sum test).

3.1.11 Influence of Histone Modifications on TF Target Gene Regulation

The high overlap of TFBS with histone modified sites prompted the question whether this co-occurrence has an influence on transcript level. Therefore, transcripts were categorized into TF targets with and without a histone modified site (Figure 27). Using Wilcoxon rank sum test ($p < 0.05$) to compare the mean expression level of transcripts bound solely by the individual TFs and those bound by the TF and any histone modified site, it was investigated whether the presence of histone marks had an influence on the expression level of target genes. A strong positive dependency was found for three of the four TFs, namely, Gata4, Nkx2.5 and Srf, while Mef2a bound target genes were higher expressed independent of any of the measured histone modifications.

To analyze the extent to which each histone modification contributes to this increase, the targets showing histone modified sites were further grouped according to the individual modifications. Again Wilcoxon rank sum tests were applied to determine significant differences, now against genes showing only TF binding without an additional histone modification (Figure 28). Interestingly, although each of the TFs and each of the histone modifications were individually found to be associated with higher transcript levels, the effect of the combinations differed for each TF. In case of Nkx2.5 (Figure 28 B) all four histone modifications were associated to higher transcript levels, each to almost the same extent. For Gata4 (Figure 28 A), each of the four histone modifications was significantly associated to higher transcript levels except H3K4me3, however, the elevation was most significant for H3ac. In case of Srf (Figure 28 D), this effect was even more prominent with a high influence of H3ac and much less influence of H3K4me2, H3K4me3 and H4ac. In line with the analysis of any histone modification, no single histone modification was found that showed elevated transcription levels of Mef2a targets compared to binding without any histone modification (Figure 28 C).

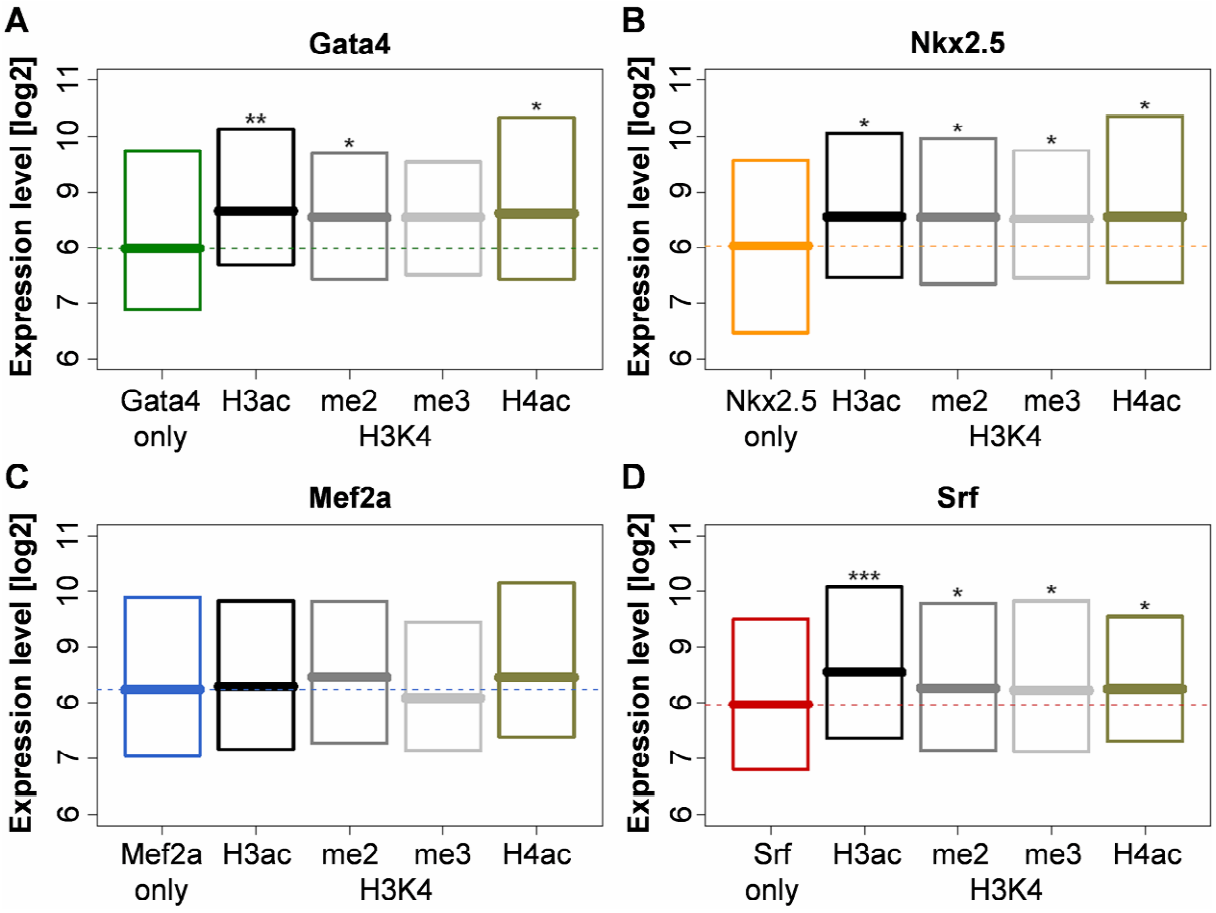

Figure 28: Influence of individual histone modifications on target gene expression
Binding sites for each TF were categorized into groups depending on co-occurrence with histone modifications in ChIP-chip. Transcripts without co-occurring histone modified sites were used as a reference (TF only). Stars on top of each box indicate significant difference from the reference group. Stars between two boxes indicate a significant difference between the two groups. Significance levels are depicted as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$ (Wilcoxon rank sum test).

Because a single gene typically shows multiple histone modifications and will therefore belong to more than one of the defined groups, analyzing the difference in target gene expression as described will result in a biased estimate for the influence of each histone modification as. As for Srf targets the influence of H3ac was found to be much more prominent than any other histone modification, it was tested if the low activating effect seen for the other histone modifications was merely due to their co-occurrence with H3ac. Therefore, the influence of each modification was estimated using the linear model

$$Y = \beta_0 + \beta_{\text{H3ac}} + \beta_{\text{H3K4me2}} + \beta_{\text{H3K4me3}} + \beta_{\text{H4ac}} + \varepsilon \quad ,$$

where the observed variable Y was the gene expression of an Srf target, the predictive variables refer to the individual histone modifications and β_0 models the expression baseline. As each of the predictors is categorical (each histone modification is either absent or present) this linear model is equivalent to an ANOVA. After least square estimation of the coefficients, it was tested using F statistics whether the full model and each of the individual coefficients was significantly different from zero and therefore had a significant influence on the expression of Srf targets.

	Intercept	H3ac	H3K4me2	H3K4me3	H4ac
Estimate	8.376	1.035	-0.106	-0.397	-0.223
p-value	$< 2x10^{-16}$	$4.4x10^{-6}$	0.633	0.092	0.237

Table 12: ANOVA estimates and p-values
The p-values are based on *F* statistics and reflect the significance of the estimates being different from zero.

Again, the resulting linear model showed an overall significant dependency of Srf target gene expression on accompanying histone modifications ($p = 7.301x10^{-5}$). However, when the individual estimates for each histone modification were considered only the influence of H3ac remained significant with a p-value of $4.4x10^{-6}$ (Table 12). This underlines the influence of co-occurring H3ac marks on Srf target gene expression and revealed that the activating effect seen for the other histone modifications was only due to their co-occurrence with H3ac.

3.1.12 Read Mapping and Peak Calling for the ChIP-seq Data

To confirm and further investigate the impact of H3ac on Srf target gene expression, genome-wide ChIP followed by next-generation sequencing (ChIP-seq, section 2.2.3) was performed using antibodies against both Srf and H3ac in HL-1 cells. The sequencing of the individual ChIP experiments resulted in 6,967,318 reads for Srf and 8,364,328 reads for H3ac. These were mapped to the mouse reference genome (NCBI m37) using the read mapping tool RazerS[163] as described in section 2.3.6 allowing at most two mismatches and no indels. The mapping resulted in 4,543,634 (65.2%) of mappable reads for Srf and 6,141,144 reads (73.4%) for H3ac indicating good experimental qualities. The error distribution of reads for both experiments is depicted in Table 13 and Figure 29. After the mapping, reads that were assigned to multiple genomic regions were filtered out.

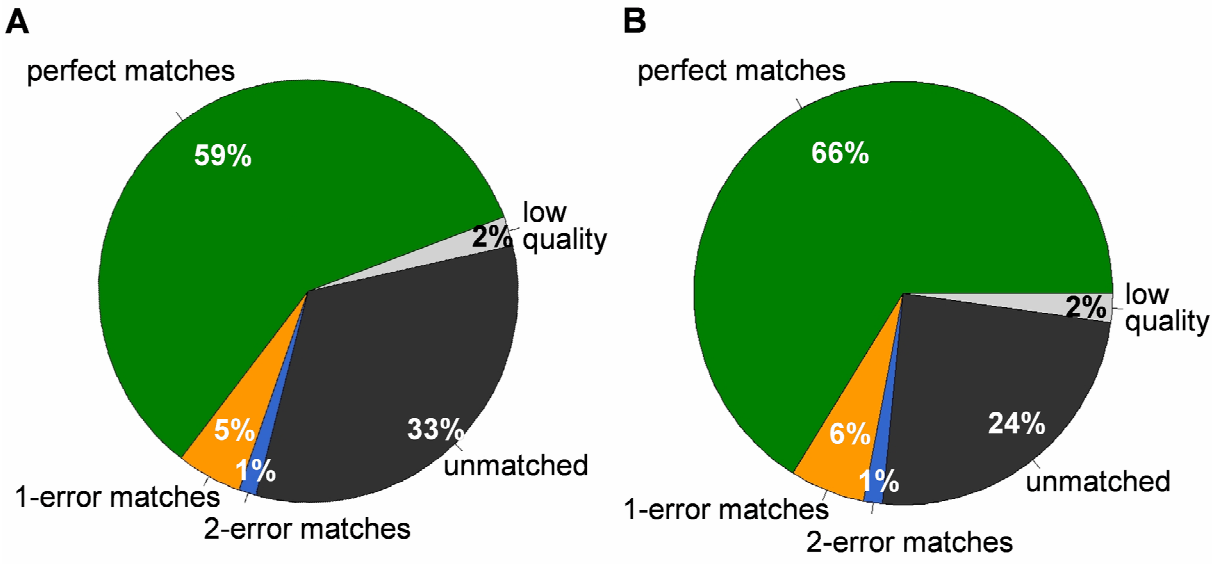

Figure 29: Distribution of ChIP-seq read matches
(A) Srf and **(B)** H3ac ChIP-seq read were mapped to the mouse reference genome using RazerS. Reads that matched without any error (perfect matches, green), with one error (1-error matches, orange) and with two errors (2-error matches, blue) were retained, while matches that could not be mapped to the mouse genome (unmatched, black) or were of low quality (gray) were discarded.

	Srf	H3ac
Total number of sequenced read	6,967,318	8,364,328
Number of low quality read	156,845 (2%)	183,557 (2%)
Number of perfect matches	4,096,439 (59%)	5,531,016 (66%)
Number of 1-error matches	350,057 (5%)	487,420 (6%)
Number of 2-error matches	97,138 (1%)	122,708 (1%)
Number of unmatched read	2,266,839 (33%)	2,039,627 (24%)
Number of called peaks	2,190	10,486

Table 13: Number of ChIP-seq read matches and called peaks
Reads that matched without any error (perfect matches), with one error (1-error matches) and with two errors (2-error matches) were retained while matches that could not be mapped to the mouse genome (unmatched) or were of low quality were discarded before the peak calling. Percentages are computed in respect to the total number of reads.

As ChIP-seq data differs greatly from ChIP-chip data (probe intensities versus read counts) a different peak calling approach (implemented in the CisGenome software package; section 2.3.6) was applied. For the Srf ChIP-seq data the CisGenome software was used with a window size of 100 bp, a step size of 25 bp for the sliding and a minimal read count level of 10 resulting in an estimated FDR of 1.6%. As histone enriched sites were shown to be broader than transcription factor peaks,[16] a window size of 250 bp, a step size of 50 bp for the sliding and a minimal read count level of 10 was used for the H3ac ChIP-seq data, resulting in an estimated FDR of 4.7%. After the peak calling procedure, boundary refinement (Srf and H3ac) and single-strand filtering (only H3ac) were applied. After manual inspection of individual peaks, no single-strand filtering for the Srf ChIP-seq data was performed. Finally, the ChIP-seq approach identified 2,190 and 10,486 peaks for Srf and H3ac, respectively. These were associated to 1,902 and 10,689 genes, respectively, using the same criteria as for the ChIP-chip data.

3.1.13 Analysis of ChIP-seq Results and Comparison to ChIP-chip

As the ChIP-seq and ChIP-chip approach both aim to measure the same enriched binding sites but use different techniques, it was interesting how high the overlap between these two techniques would be.

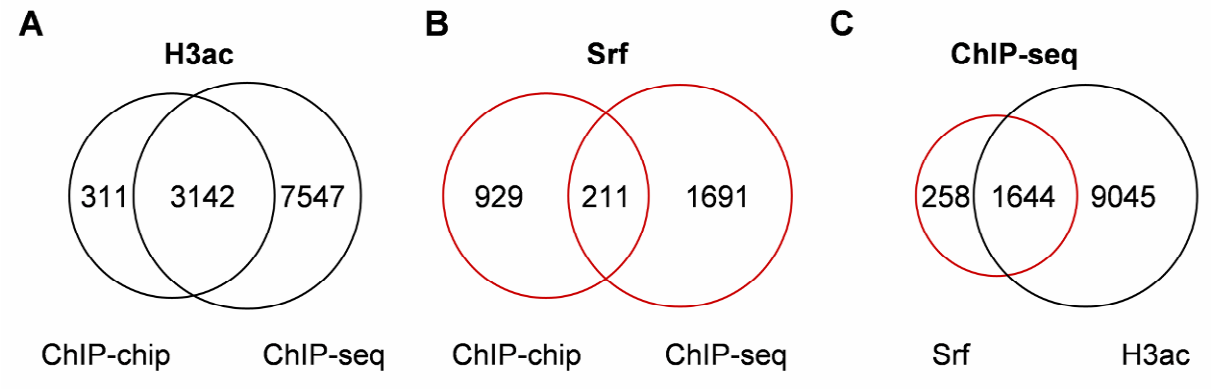

Figure 30: Target genes of Srf and H3ac in ChIP-chip and ChIP-seq
(A+B) Overlap between genes associated to **(A)** H3ac and **(B)** Srf peaks in ChIP-chip compared to ChIP-seq. Note that the number of analyzed genes was much higher for ChIP-seq. **(C)** Overlap between Srf and H3ac target genes in ChIP-seq.

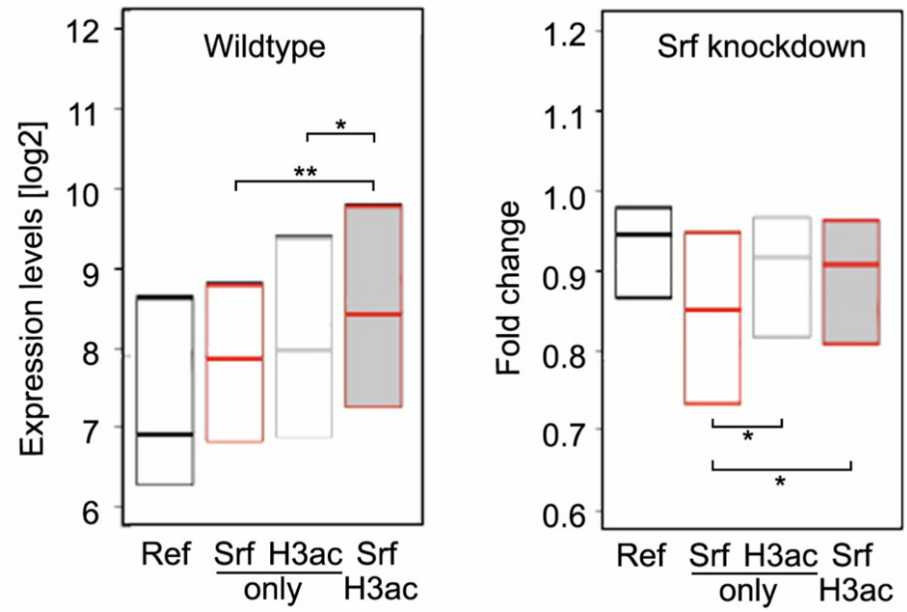

Figure 31: Confirmation of H3ac depending expression of Srf targets by ChIP-seq
(A) Boxplots of expression levels of transcripts grouped according to H3ac and/or Srf binding close to the transcriptional start site (TSS < 1.5kb). **(B)** Boxplots of fold changes relative to siNon of down regulated transcripts after Srf knockdown grouped according to H3ac and/or Srf binding close to the transcriptional start site (TSS < 1.5 kb). **(A+B)** The resulting p-values are indicated as follows: **: $p < 0.01$ and *: $p < 0.05$. Genes with neither H3ac nor Srf sites were used as reference.

From 3,453 genes that were defined to be associated to histone 3 acetylation peaks in ChIP-chip, 91% overlapped with the ChIP-seq data (Figure 30 A). However, for the 1,150 genes associated to ChIP-chip Srf peaks the overlap was only 19% (Figure 30 B). This low overlap hits a current debate in the ChIP community. A comparison for NRSF peaks conducted by Ji *et al.*[94] showed that only 22% of their ChIP-chip peaks overlapped with ChIP-seq peaks but that the overlapping peaks had a much higher number of observed motifs than those that occurred only in ChIP-chip or ChIP-seq. Schones et al.,[93] Alekseyenko et al.[219] and Choi et al.[220] have addressed this problem further. Summarizing their results, the two technologies show a clearly different behavior in terms of sensitivity and specificity with potentially additive information content. While ChIP-seq peaks tend to form regions that are much sharper than those in ChIP-chip due to its superior resolution, ChIP-chip peaks might additionally cover binding events with more moderate significance. This would fit to the observation, that the overlap of Srf peaks was much smaller than that of H3ac peaks, as the latter exhibit a much stronger signal in the ChIP experiment. Finally most (86%) of the Srf target genes were found to have an additional H3ac modified site (Figure 30 C).

Despite the differences found between Srf target genes in ChIP-chip versus ChIP-seq, analyzing the ChIP-seq data in the same way as the ChIP-chip data, a similar synergistic effect of H3ac and Srf binding was found when compared to non-bound genes or genes solely bound by either of the two (Figure 31 A). The influence of histone 3 acetylation marks was further substantiated by integrating the ChIP-seq results with the siRNA knockdown data of Srf in HL-1 cells. In accordance to its mainly activating function in wildtype cells, a significant decrease was found in the expression levels of genes bound by Srf without any additional H3ac marks. However, this decrease was absent in genes that were additionally marked by H3ac pointing to a buffering effect of H3ac on Srf target gene expression after reduction of the Srf protein (Figure 31 B).

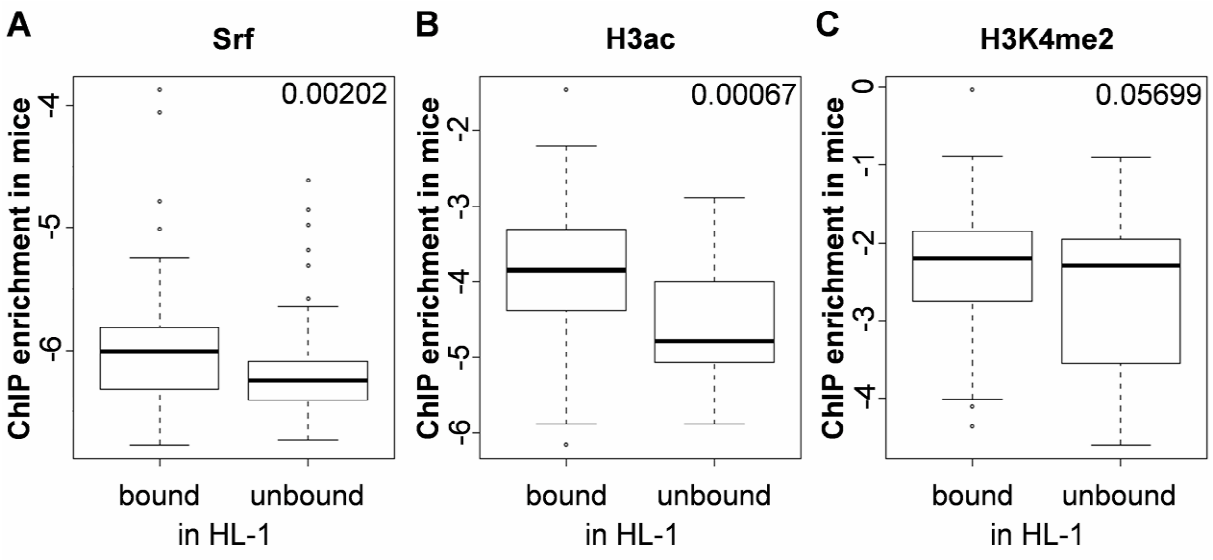

Figure 32: Comparison ChIP results gathered in mouse hearts to HL-1 cells
Boxplots comparing ChIP qPCR enrichment in mouse hearts with regions showing enrichment of Srf, H3ac or H3K4me2 using ChIP-chip/seq in HL-1 cells. For each factor the regions that were bound in HL-1 cells also had a higher average enrichment. T-test p-values for the difference in mean are indicated in the upper right corner.

3.2 Time Series Data of Histone Modifications and Transcription Factor Binding During Cardiac Maturation

In an attempt to confirm and further investigate the results gathered in cell culture, Srf and H3ac binding and their influence on gene expression was studied in mouse hearts in a time series during cardiac maturation at thee developmental stages E18.5, P0.5 and P4.5. In addition, two further contributing factors, namely the histone acetyltransferase p300 and the histone tail modification H3K4me2, were measured. P300 was previously suggested to be recruited by Srf,[62,221] and its acetylation efficiency was shown to be correlated to the presence of H3K4 methylation.[222]

Due to the low amount of tissue that can be gathered from mouse hearts, qPCR was used to measure the ChIP enrichment of the four factors in each stage. However, different to *e.g.* next-generation sequencing, qPCR requires an *a priori* definition of genomic regions that should be analyzed, like TFBS or histone modified sites. This selection of regions with a likely regulatory background, which is described in section 2.2.4, was based on the results from the respective ChIP data. In total, 191 regions were selected and ChIP followed by qPCR was performed for every single region with samples of mouse hearts of each individual time point (E18.5, P0.5 and P4.5) and each measured TF and histone modification (Srf, p300, H3ac and H3K4me2,) as well as Input DNA in triplicates. Finally, the measured Input enrichment was used to calculate relative DNA amounts.

3.2.1 Preliminary Analysis

As a proof of principle and to check if the heart tissue would reveal ChIP enrichments similar HL-1 cells, these were compared to P0.5 results (Figure 32). P0.5 was used as HL-1 cells have been originated from postnatal right atrium heart cells. Using t-tests it was shown that the mean enrichments of regions bound by H3ac, H3K4me2 and Srf in HL-1 were significantly elevated also in mouse hearts. This shows that the ChIP-qPCR experiments resulted in valuable results and the heart tissue

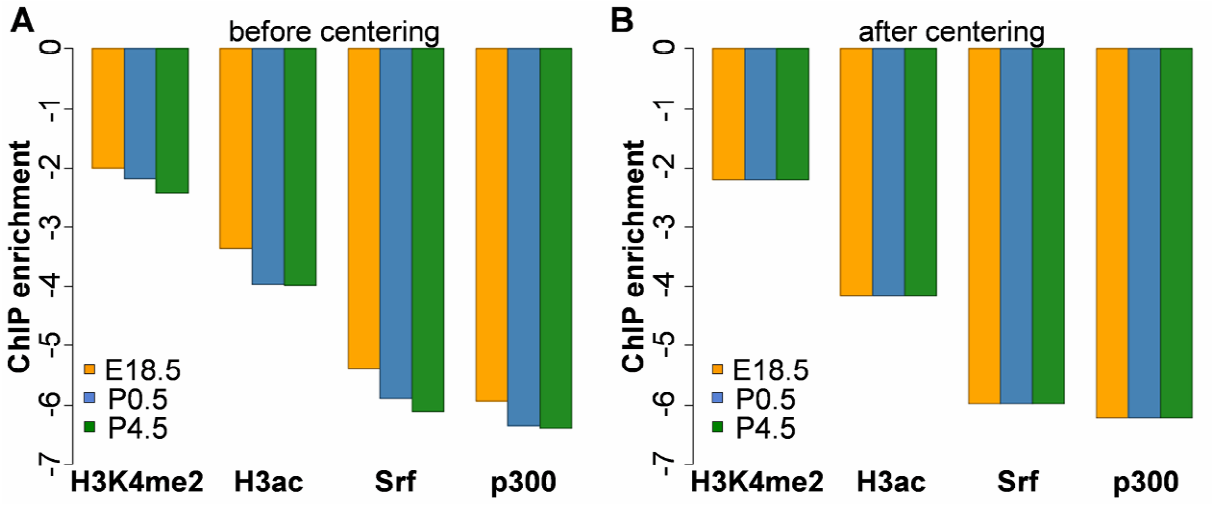

Figure 33: Barplot of average enrichment over all regions for every factor in every stage
(A) Unnormalized ChIP enrichments showing distinct trends between the individual stages. **(B)** Measurements after linear shifting which removes the trends.

was sufficiently close to the previously studied HL-1 cells. P300 was not measured in HL-1 and could therefore not be compared.

To study the dynamics of transcriptional regulation, changes between the individual stages were analyzed. Consequently, the average enrichment for each factor in each stage was compared and found to be different between the three stages (Figure 33 A and Figure 34 A). While these differences closely recapitulated findings from a simultaneously performed protein analysis, they would also result in general trends in binding strength changes between the stages that would lead to a superficially high correlation between changes in the individual factors in any subsequent analysis. To eliminate these general trends a linear shift was introduced for each single region and each single factor resulting in the same average enrichment in each stage for each (Figure 33 B and Figure 34 B).

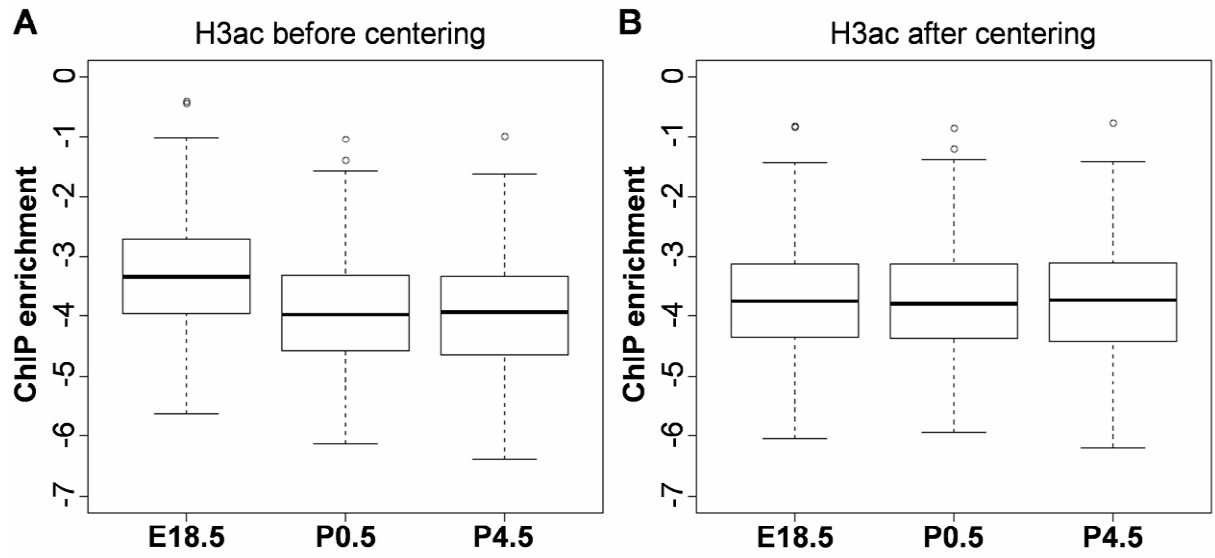

Figure 34: Boxplot of average enrichment over all regions for H3ac in every stage
(A) Measurements after ΔCP normalization showing a distinct trend between the individual stages. **(B)** Measurements after linear shifting which removes the trend.

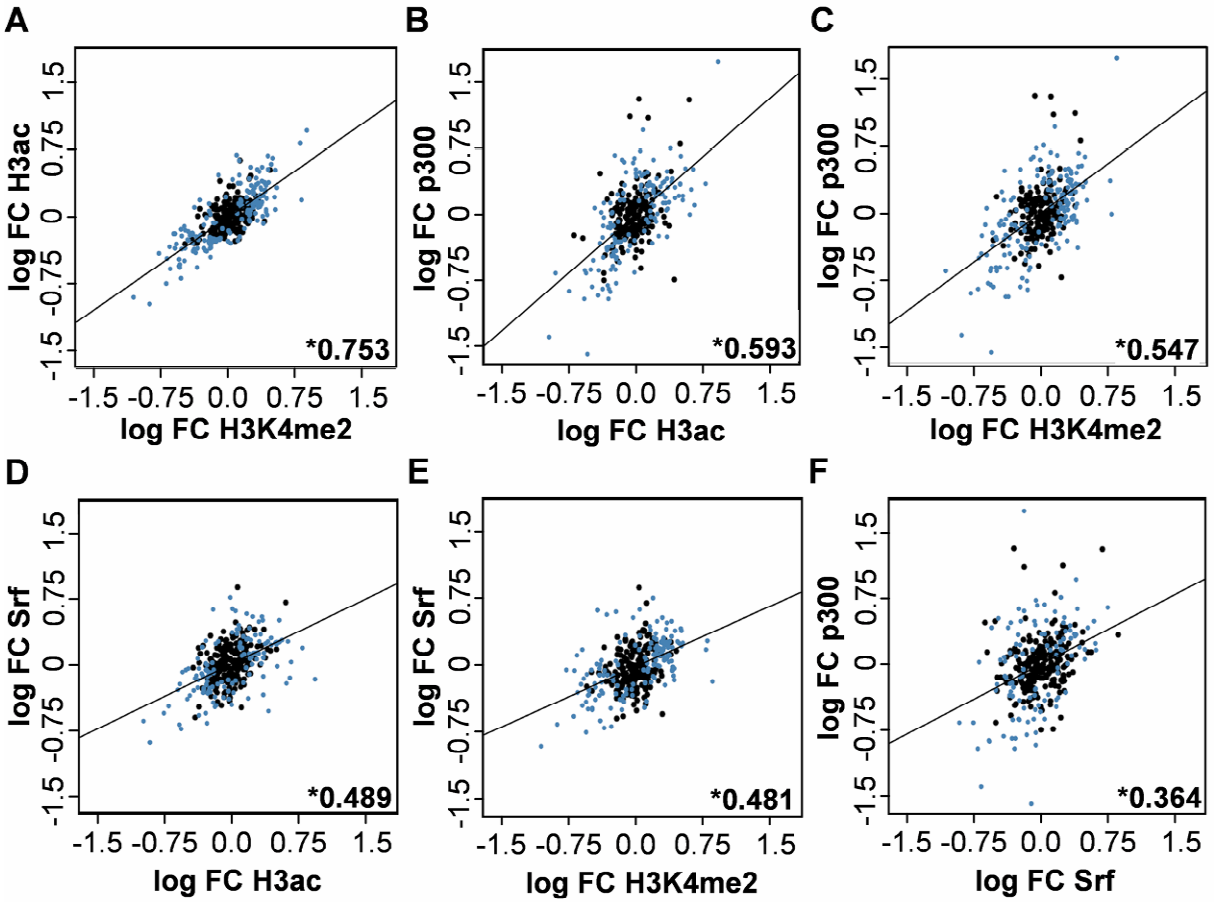

Figure 35: Scatter plot of fold changes between the measured factors (combined time points)
Blue dots represent measurements with significant change for at least one of the factors according to t-test. Lines represent the best linear fit for all measurements. Pearson correlation coefficient over all measurements is indicated in the lower right corner. Significance (p-value < 0.001) is indicated by *.

3.2.2 Analysis of Correlated Binding Changes

After combining replicates using the arithmetic mean, log fold changes between mean enrichments in the two consecutive stages (P0.5/E18.5 and P4.5/P0.5) were calculated. In addition, each fold change was categorized as up, down or unchanged according to a t-test using a significance level of 0.05.

As a first analysis, scatter plots were used to visualize the level of pairwise correlation between the fold changes of individual factors incorporating both time points (Figure 35). High to modest correlations were found for all the measured factors, with Pearson correlation coefficients ranging from 0.75 to 0.36. To evaluate the statistical significance of the observed correlation coefficients empirical p-values were derived using random experiments. As a null model, measurements for one factor and time point were randomly assigned to all regions without replacement and fold changes were again computed between successive timepoints. This process was repeated 20,000 times storing all pairwise correlation coefficients between individual factors. An empirical p-value was derived by counting the number of tries were the random correlation coefficient exceeded the true coefficient. Figure 36 shows the distribution of Pearson correlation coefficients observed in the random experiment in comparison to the correlation coefficients found in the real data. Given the large sample size (almost 400 for the analysis of combined timepoints) random correlation coefficients scatter between -0.2 and 0.2. Applying a significance level of at least 0.001 all of the analyzed pairings were found to show significant correlation, however with different significance levels.

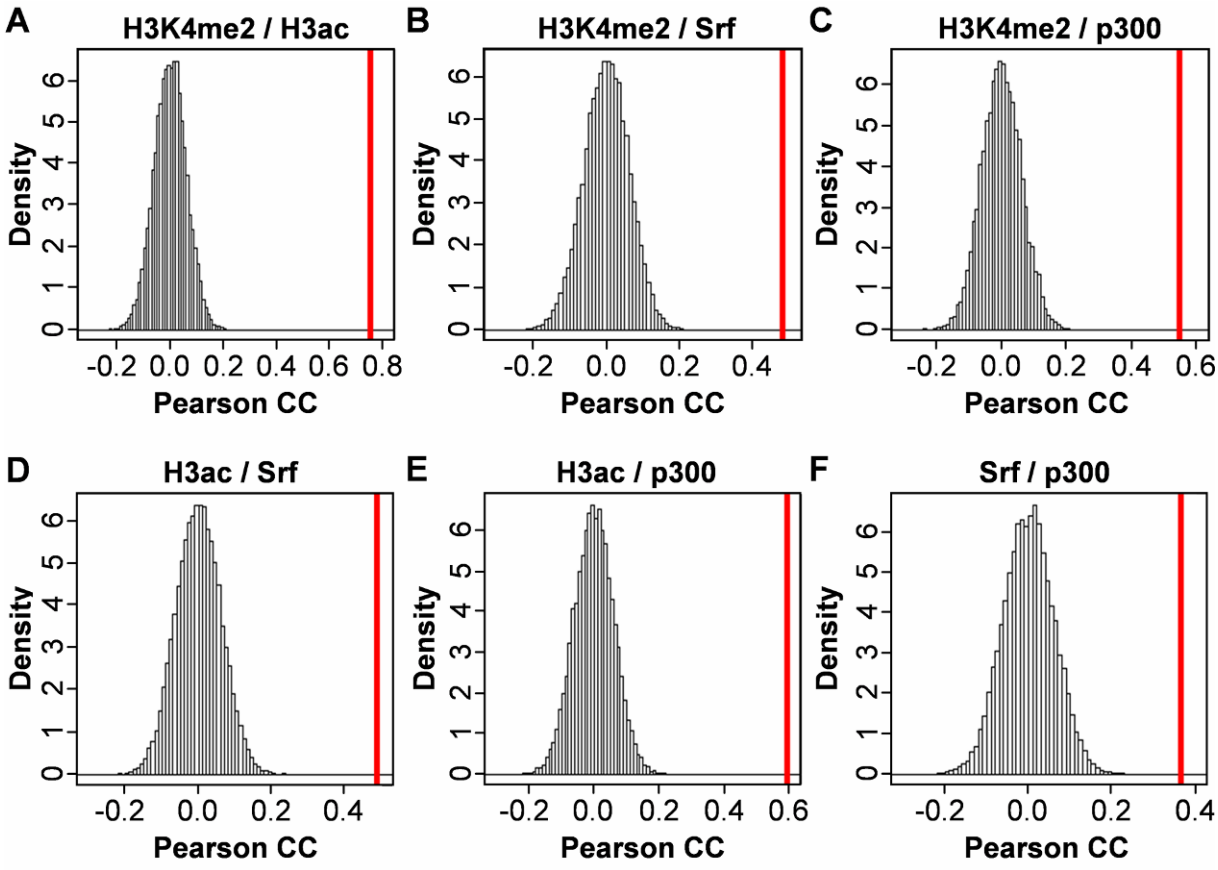

Figure 36: Distribution of Pearson correlation coefficients for real and random data
Histograms showing the distribution of Pearson correlation coefficients resulting from the random experiments. The correlation coefficients observed in the real data are indicated by a red vertical line. Data shown for combined time points.

The highest correlation coefficient of 0.75 was found between the changes in histone 3 acetylation and histone 3 lysine 4 dimethylation. This is in accordance to the high correlation of absolute H3ac and H3K4me2 levels found in previous studies from us[16] and others.[223] In line with this, a recent study performed by Wang *et al.*[107,222] suggested that H3K4 methylation might directly facilitate the histone acetylation events. The second highest correlation coefficient (0.59) was found between changes in H3ac enrichment and p300 followed by p300 and H3K4me2 (0.54). P300 is a known histone acetyltransferase and transcriptional co-activator[224] and was found to reside in enhancer and promoter regions.[225,226] While the correlation between changes in p300 and acetylation level was highly expected given the function of p300, the correlation between p300 and methylation level is of high interest. A possible explanations for this correlation supposed by Pray-Grant *et al.*[227] and Wysocka *et al.*[228] is an initial opening of chromatin through recruitment of ATP-dependent chromatin remodeling complexes initiated by the presence of histone methylation which then allows p300 to bind. Another possible mechanism is the direct recognition of methylated sites by histone acetylation complexes (including p300) as suggested by Martin *et al.*[229]

In addition, significant correlation was found between Srf and H3ac (0.48) as well as Srf and H3K4me2 (0.481). A possible mechanistic link between these correlated changes in enrichment and binding strength, which was previously proposed for smooth muscle cells, is the stabilization of Srf binding to the CArG-box DNA motif via Myocardin or some Myocardin related factors that directly bind methylated histones (reviewed by McDonald *et al.*[230]). Srf and Myocardin further recruit other

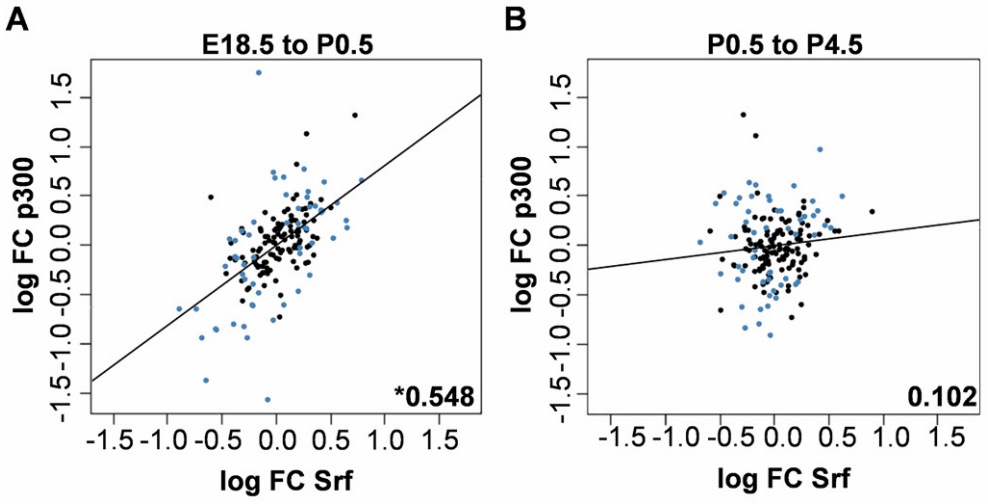

Figure 37: Correlation between Srf and p300 is dependent on the time point
Scatter plot of fold changes for p300 and Srf between **(A)** E18.5 and P0.5 and **(B)** P0.5 and P4.5. Blue dots represent measurements with significant changes for at least one of the TFs. Lines represent the best linear fit for all measurements. Pearson correlation coefficient over all measurements is indicated in the lower right corner. Significance (empirical p-value < 0.001) is indicated by *.

transcription factors including p300 which in turn lead to higher acetylation levels. Consistent with this model, positive correlation between changes in Srf and p300 could be detected, however with a very modest correlation coefficient of 0.36. Looking at the distribution of random correlation coefficients, the correlation between Srf and p300 was significant, though, the fact that the observed correlation between these two TFs is much weaker than between any of the other measured regulatory factors and especially between Srf and H3ac suggests the existence of additional mechanisms of Srf-triggered acetylation which are independent from p300 binding.

While the initial analysis of correlation between changes in the individual factors was based on the combination of both fold changes, each fold change was further analyzed individually. In general, correlation coefficients were higher for changes between E18.5 and P0.5 than between P0.5 and P4.5 while still remaining significant. This finding can be explained by the fact that fold changes between P0.5 and P4.5 were in general lower than between E18.5 and P0.5 pointing to a higher similarity between the two postnatal stages than between the prenatal and the postnatal stage. This potentially causes a higher influence of experimental noise on the correlation.

As an exception, the correlation between changes in Srf and p300 was almost completely lost between P0.5 and P4.5 (Figure 37) with a very low correlation coefficient of 0.1 and a non-significant empirical p-value. An explanation independent from the similarity of the two postembryonic stages, would be a diminished regulatory association between Srf and p300 after birth. This could be linked to the highly reduced amount of activating histone modifications that was found in the protein level analysis correlating to a smaller number of transcribed genes in postembryonic cardiomyocytes.

3.2.3 Single-Factor Qualitative Models (ANOVA)

To study the inter-dependency between H3ac, Srf and p300 in more detail, linear modeling was applied. Based on the Srf target gene activation model by McDonald *et al.*,[231] which places the acetylation as the last step in the regulatory chain, linear models of the form

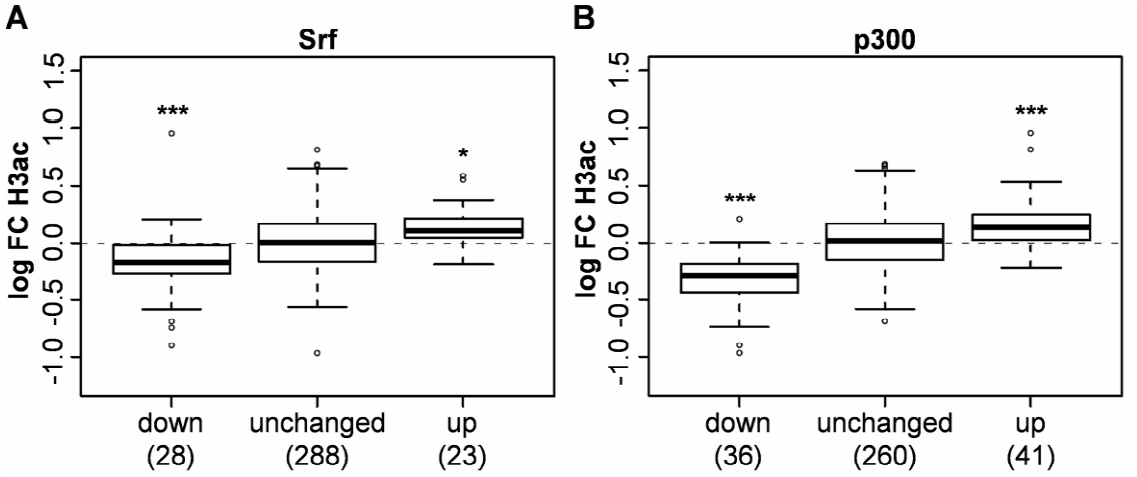

Figure 38: Boxplots for single-factor ANOVAs predicting changes in H3ac (combined time points).
Boxplots illustrating the dependence of changes in H3ac from categorized changes in **(A)** Srf and **(B)** p300. The number of regions in each group is indicated in brackets. Significance levels indicating difference from zero according to the linear model are given above each box, using the following coding: *: $p < 0.05$ **: $p < 0.01$ ***: $p < 0.001$.

$$Y_{H3ac} = \beta_0 + \beta_{Srf} + \beta_{p300} + \varepsilon$$

were defined, which predict fold changes of H3ac enrichment by fold changes of p300 and Srf enrichment. The H3K4me2 data was excluded as this was shown to be highly correlated with the dependant variable H3ac and could therefore obscure the models.

First, two single-factor ANOVAs were performed using categorized changes in either p300 or Srf as the predictor variable (Figure 38 A and B). After estimating coefficients using least square estimation,

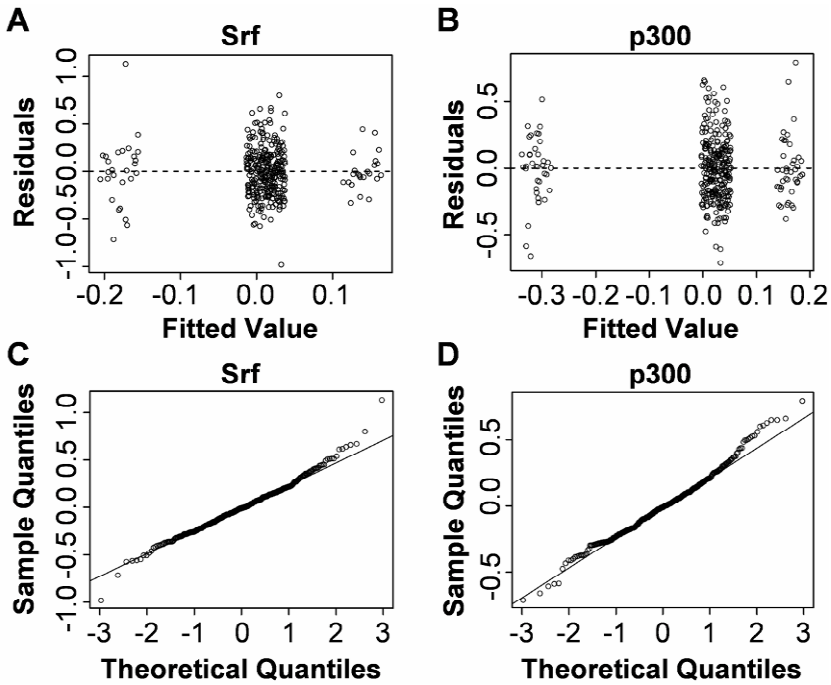

Figure 39: Quality check of single-factor ANOVA (combined time points)
Residuals against fitted values for **(A)** Srf and **(B)** p300. Q-Q normal plot for **(C)** Srf and **(D)** p300.

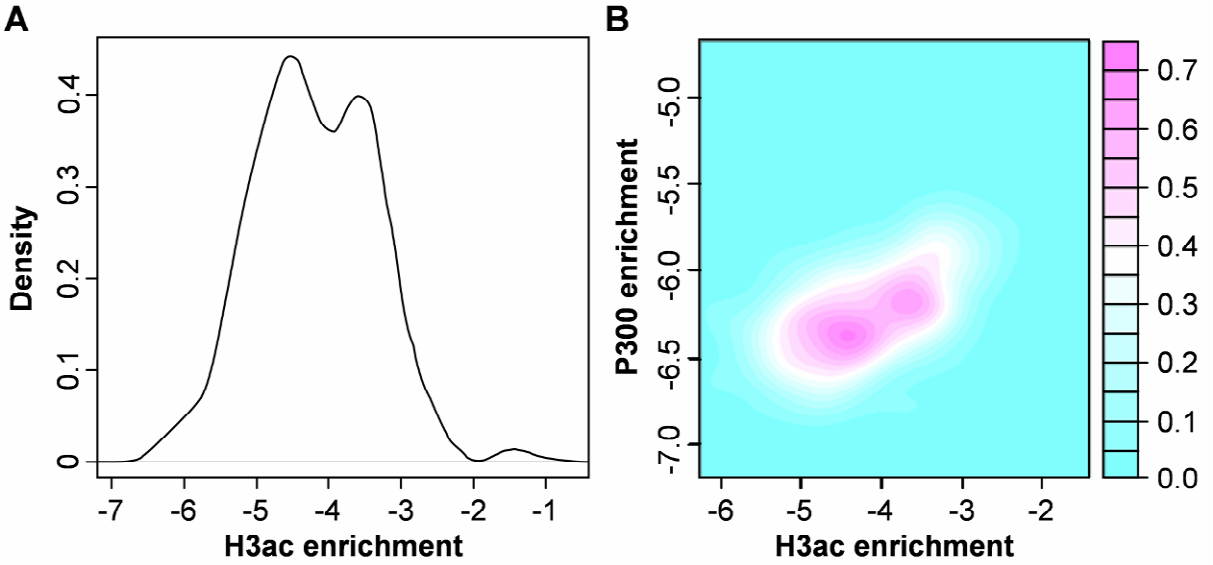

Figure 40: Distribution of enrichment levels for regions with unchanged H3ac (combined time points)
Mean enrichment in E18.5 and P0.5 were used as measurements for the absolute enrichment between the timepoints **(A)** One-dimensional density plot for histone 3 acetylation enrichment levels using a bandwidth of 0.253 and n=227 observations. **(B)** Two-dimensional density plot for H3ac and p300 enrichment levels. The height is indicated using the color key on the right. The two-dimensional density was computed using the kde2d function from R's MASS package.[232]

it was tested with *F* statistics whether each of the individual factors was significantly different from zero and has therefore a significant influence on changes in H3ac enrichment. As before, randomized experiments were used to derive empirical p-values for the individual models, now using the *coefficient of determination* R^2 as the measure for the goodness of fit for each model. An empirical p-value was determined by the number of random models which yielded a higher R^2 than the original model.

In line with the results from the correlation analysis, both models indicated a significant dependency of H3ac on the individual TF with an empirical p-value < $5x10^{-5}$. Further, the estimated changes in H3ac enrichment level for Srf/p300 up and down regulated regions were significantly higher and lower than zero, respectively, while the estimate for regions without a significant change, which were used as a control group, did not show this difference in either model. The distribution of the residuals (Figure 39 A and C) and the normal Q-Q plots (Figure 39 B and D) indicated validity of the linear model assumptions for both single-factor ANOVAs.

As the estimated average change in H3ac for those regions without any significant change in either p300 or Srf enrichment was almost zero, it was inspected if these non-changing regions had a general preference in their absolute H3ac enrichment levels. Figure 40 A shows the density plot of mean enrichment levels for H3ac over all regions. It indicates a bimodal distribution with two peaks that correlate with regions of high and low acetylation enrichment. The same two peaks were found in a two dimensional density plot for absolute H3ac and p300 enrichment levels (Figure 40 B) but not for H3ac and Srf. However, in both plots the two peaks are not well separated and many regions show H3ac and p300 enrichment levels between these two extremes making a simple interpretation that is only based on these two stages difficult.

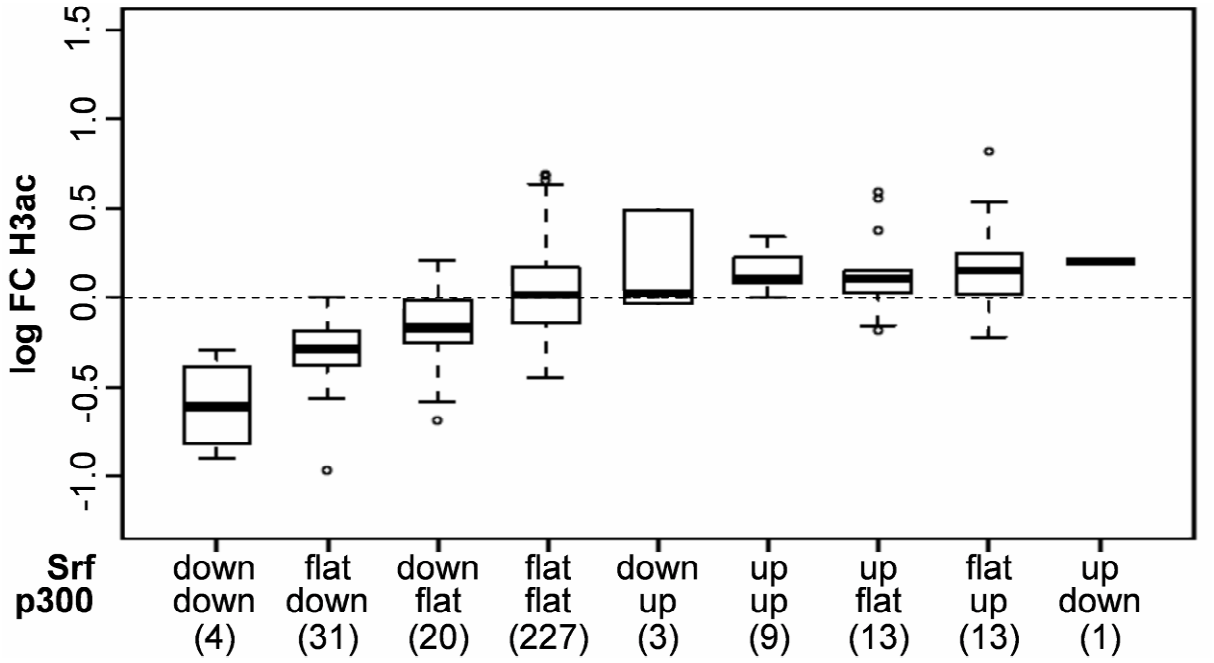

Figure 41: Boxplot for two-factor ANOVA predicting changes of H3ac from changes in Srf and p300 (combined time points)
Boxes are ordered according to the median of the appropriate group. The number of regions with significant changes are indicates in brackets below the group name. Unchanged regions are indicated by the term 'flat'.

3.2.4 Two-Factor Qualitative Models (ANOVA)

After proving the dependency on each factor individually, their regulatory interaction was analyzed in more detail. To do so, a two-factor ANOVA was performed using categorized changes in both Srf as well as p300 as predictors for changes in H3ac (Figure 41). Again, the model indicated significant dependency with an empirical p-value < $5x10^{-5}$. Like in the single-factor ANOVA, the estimates for the influence of up and down regulation of Srf and p300, which can be found in Table 14, were significantly different from zero, with the exception of up regulated Srf regions. The fact that changes in both p300 and Srf have a significant influence on H3ac changes rather than p300 alone further substantiates the assumption of an additional mechanism of histone 3 acetylation triggered by Srf which is independent of p300. Again, the distribution of the residuals (Figure 42 A) and the normal Q-Q plots (Figure 42 B) indicated validity of the linear model assumptions for both ANOVAs.

3.2.5 Quantitative Models

After using categorized changes for the modeling, it was tested if H3ac levels were also quantitatively dependent on Srf and p300 levels. In other words, whether stronger changes in any of these two factors lead to stronger changes in histone 3 acetylation as suggested by the high pairwise correlations

	Intercept	p300 up	p300 down	Srf up	Srf down
Estimate	0.03483	0.12557	-0.32705	0.06896	-0.18069
p-value	$2.23x10^{-2}$	$1.9x10^{-3}$	$4.56x10^{-14}$	$1.84x10^{-1}$	$1.42x10^{-4}$

Table 14: Estimates and p-values of two-factor ANOVA
The p-values are based on *F* statistic and reflect the significance of the estimates difference from zero. All estimates are in log space.

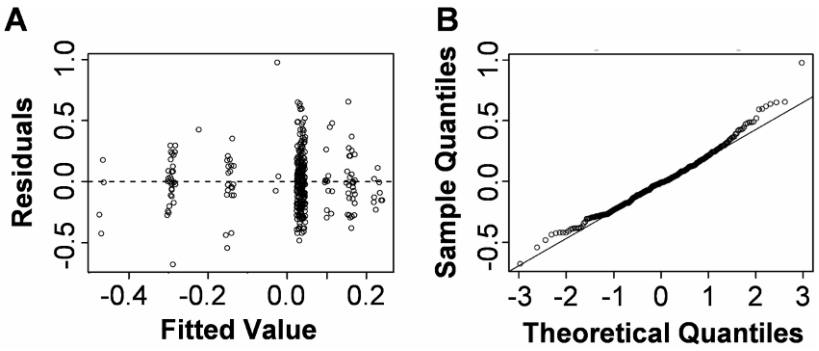

Figure 42: Quality check of two-factor ANOVA (combined time points)
Residuals against fitted values for **(A)** Srf + p300 and **(B)** Srf + p300 + interaction term. Q-Q normal plots for **(C)** Srf + p300 and **(D)** Srf + p300 + interaction term.

and which of the two factors was more influential. Therefore, a linear regression model was built again using H3ac fold changes as the dependant variable and Srf and p300 fold changes as the predictive variables.

According to both models, absolute changes in H3ac are positively correlated with changes in Srf as well as p300 and to a roughly equal extent according to the high similarity in their estimated coefficients (Table 15). Like for the ANOVA models, the empirically derived p-value for the quantitative model was significant ($< 5x10^{-5}$) and the distribution of the residuals (Figure 43 A) and the normal Q-Q plots (Figure 43 C) indicated validity of the linear model assumptions for both linear regression models. In summary, the analysis of the quantitative model revealed that changes in histone 3 acetylation level were not only qualitatively dependent on changes in Srf and p300 binding level but also had a quantitative dependence with correlations between predicted and observed changes around 0.67 (Figure 43 B). While far from perfect, these correlations demonstrate that changes in the two factors Srf and p300 already drive an appreciable fraction of the measured change in histone 3

	Intercept	**Srf**	**p300**
Estimate	0.00622	0.31083	0.32681
p-value	$5.68x10^{-1}$	$5.25x10^{-12}$	$1.12x10^{-23}$

Table 15: Estimates and p-values of quantitative linear
The p-values are based on *F* statistic and reflect the significance of the estimated coefficients being different from zero. All estimates are in log space.

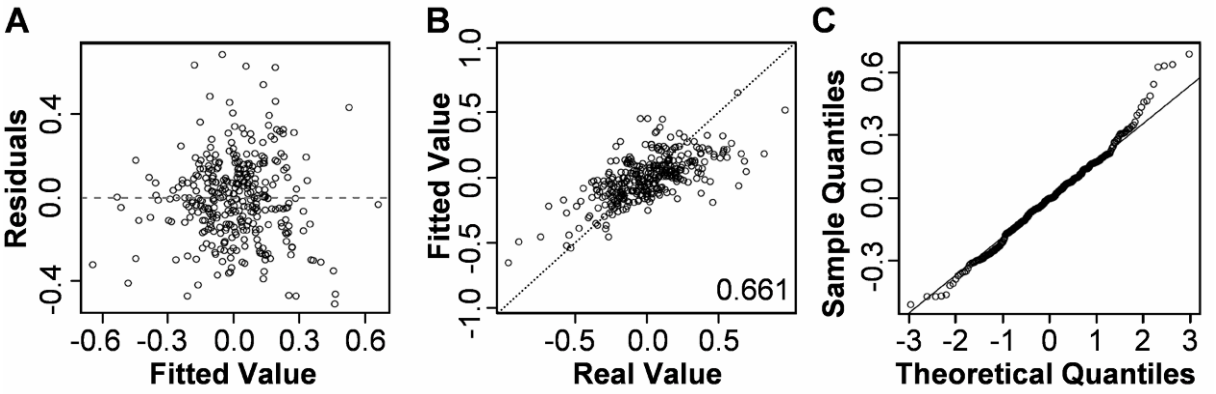

Figure 43: Quality check of quantitative model (combined time points)
(A) Residuals against fitted values for the linear model. **(B)** Model fit against measured values for the linear model **(C)** Q-Q normal plot for the linear model.

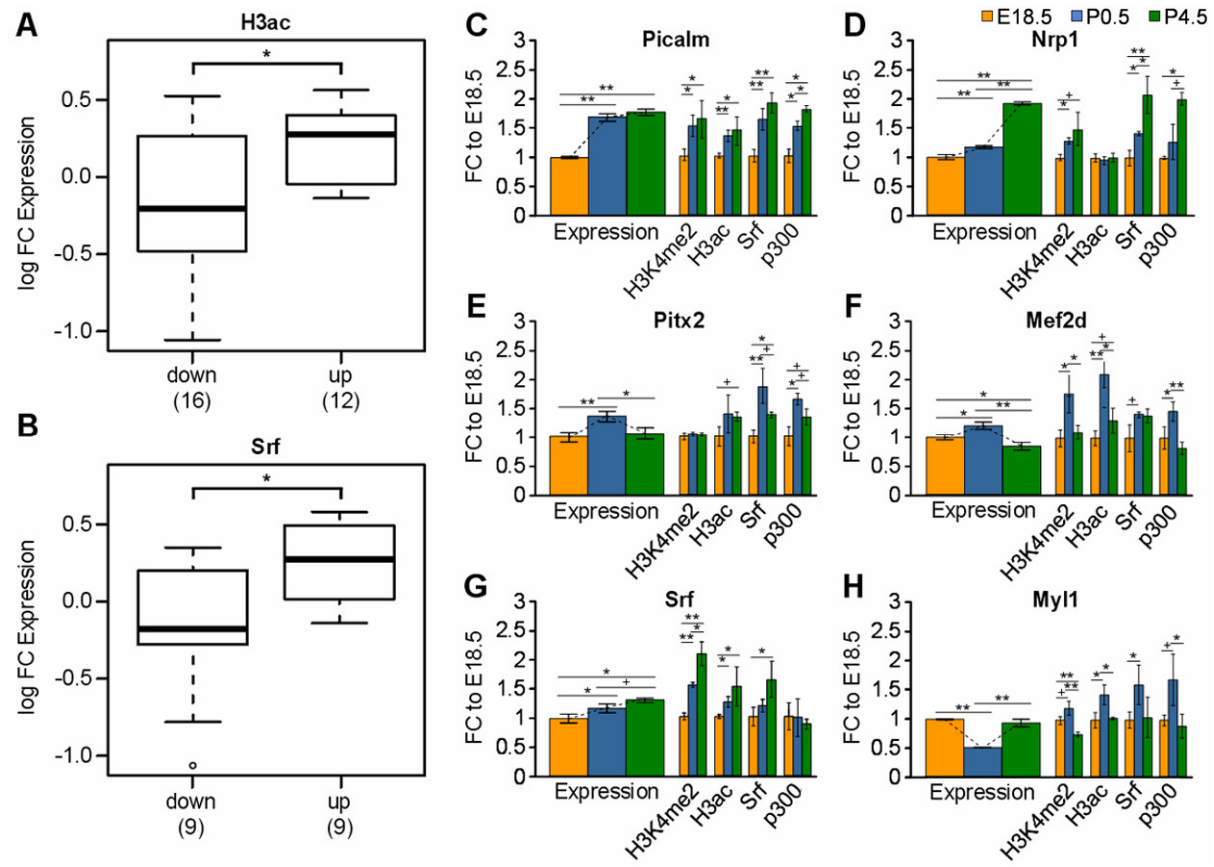

Figure 44: Consequences on gene expression
(A+B) Boxplots illustrating the dependence of changes in expression level from significant changes in **(A)** H3ac and **(B)** Srf. The number of genes belonging to each group are indicates in brackets below the group name. **(C-H)** Representative examples of variable gene regulation. **(C)** All transcription factors and histone modifications correlate with expression. **(D–G)** All transcription factors and histone modifications correlate with expression with the exception of **(D)** H3ac, **(E)** H3K4me2, **(F)** Srf and **(G)** p300. **(H)** Anti-correlation between transcription factors and histone modifications and expression. Significance levels according to t-test are depicted using the following coding: +: $p < 0.1$ *: $p < 0.05$ **: $p < 0.01$.

acetylation levels in the investigated regions.

3.2.6 Consequences on Gene Expression

In the estimation of the ANOVA and linear regression models, histone 3 acetylation levels were used as the predicted read-out of changes in Srf and p300 level. In a last step, it was analyzed if the observed changes also had a functional consequence on gene expression. Therefore, gene expression levels for a set of 44 genes associated to one or multiple analyzed regions were measured in the same three stages using quantitative real-time PCR and normalized to the housekeeping gene *Hprt*.[90] Like for the ChIP measurements, a linear shift was introduced to remove potential trends in the mean expression over all measured genes between the three individual stages and log fold changes were calculated to measure the effective change between the individual stages. For genes that had multiple regions associated, the changes for these regions were combined using the following algorithm for each factor and time point comparison individually: At first, changes were categorized into up, down and unchanged as described before. In a second step, genes that were associated to both up and down regulated regions were discarded from the analysis. Finally, genes which were associated to at least one region with a significant change were categorized accordingly and genes associated to only unchanged regions were discarded.

To determine whether changes in the individual regulatory factors lead to changes in gene expression, genes were grouped into those which had a significant up change of the respective factor and those which had a significant down change combining the two time point comparisons. Using a t-test comparing these two groups, a significant dependency for changes in gene expression on changes in H3ac enrichment ($p = 0.017$) as well as Srf binding ($p = 0.03$) was detected (Figure 44 A and B). In addition, the dependency of gene expression changes was inspected for all measured transcription factors and histone modifications manually. While the genes in general showed the proposed dependency on binding or enrichment changes, many were found that were regulated by only a subset of these (representative examples are shown in Figure 44 C-H). This finding likely indicates a high variability of combinatorial regulation between the investigated regulators. However, it should be kept in mind that only three consecutive timepoints were measured and therefore important regulatory events might have been missed.

3.3 Expression Analysis of Patient Data to Detect Disease-Associated Profiles and Predict Cardiac Regulatory Networks

To evaluate the importance of the results gathered in cell culture and mouse hearts the human cardiac regulatory network and its breakdown in heart disease was analyzed. Therefore, gene expression levels of a set of 42 genes associated to heart development and function were screened in a large number of human patients with a panel of congenital heart disease as well as a group of healthy individuals using qPCR. In addition, phenotypic information was incorporated from a clinical characterization comprising 250 features of morphological, hemodynamic and therapeutic information, which had been collected for every analyzed patient to identify specific molecular portraits for subgroups of patients with common phenotypic annotations.

3.3.1 Depicting a Phenotype Ontology

To compress the complex and partially overlapping disease characteristics present in the clinical characterization a phenotype ontology was delineated. Therefore a list of 26 key disease parameters including descriptors like '*interatrial septal defect*' and '*right ventricular dilation*' in addition to tissue type (atrium/ventricle), gender and age was compiled. The continuous parameter age was binarized into two levels, young (< 6 years) and old (> 6 years).

To define groups of patients with similar phenotypes, a hierarchical clustering approach with Euclidian distance and complete linkage (section 2.3.3) was carried out using this phenotype ontology excluding the annotations for healthy, age and gender. A manual cut-off was applied to the cluster dendrogram to finally assign patients to the eight meta-phenotype groups *TOF-I* to *TOF-IV* (Tetralogy of Fallot), *VSD* (ventricular septal defect), *TGA-PS* (transposition of the great arteries, pulmonary stenosis), *Diverse* (heterogeneous phenotypes) and *Healthy*. The heatmap of the phenotypic annotations and the resulting meta-phenotype groups can be found in Figure 45. The cluster *TOF-III e.g.* contains patients characterized by interatrial septal defects as well as stenosis and/or dilation of the main pulmonary artery in addition to the classical features of Tetralogy of Fallot, namely an interventricular septal defect, overriding aorta, right ventricular hypertrophy and right ventricular outflow stenosis (section 1.3.3).

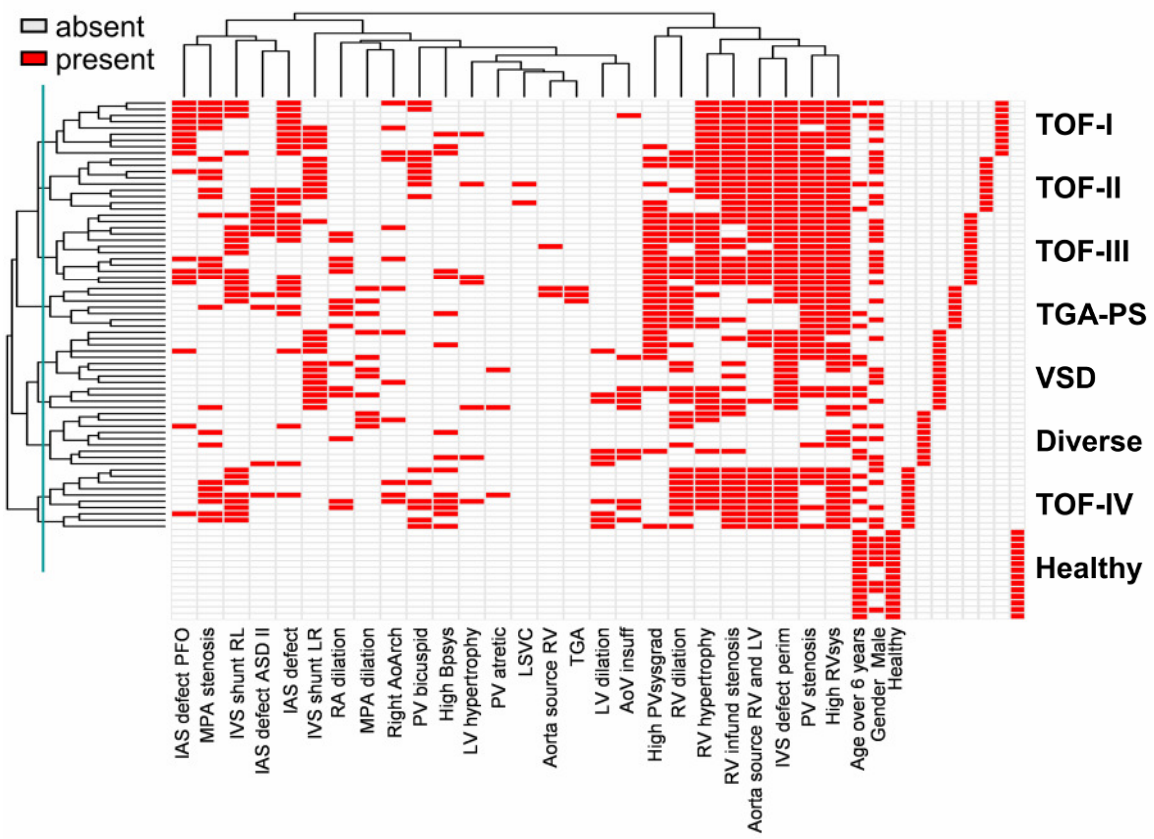

Figure 45: Phenotypes and meta-phenotypes of the analyzed patient
Shown is a hierarchical clustering of cardiac phenotype criteria and the assignment of patients with similar characteristics into meta-phenotype groups of ventricular samples. The phenotype information for gender, age and disease state is indicated. Each row represents a single heart sample. The blue line indicates the used cut-off for assignment of meta-phenotypes.

3.3.2 Preliminary Expression Data Analysis

To get an initial overview of the expression data, each expression vector was centered on zero and complete linkage hierarchical clustering was applied revealing clear difference between atrial and ventricular samples (Figure 46 A). Several of the genes displaying chamber-specific expression have already been described in studies of human and mouse myocardium. For example, *NPPA, NR2F1, MYH6, MYL7 and TAGLN* predominate in atria,[233] whereas *Irx4* and *Myl2* are restricted to ventricles.[234]

In addition, correspondence analysis was carried out. Correspondence analysis is a technique to project the high-dimensional space of the original gene/patient matrix into a lower dimensional space accounting for the main variance in the data. In that it is similar to principle component analysis, however, contrary to this projection method it accounts for the genes in the patient space as well as the patients in the gene space at the same time[235] which was more desirable for this analysis. Projecting the measured data onto two dimensions supported the tissue-specific differences and demonstrated that disease and healthy (Figure 46 B) as well as aged and young (Figure 46 C) individuals could be distinguished, implicating that the obtained data is biologically meaningful. Subsequent analyses were carried out for both cardiac tissues separately, whereof results of the ventricle are illustrated in this thesis.

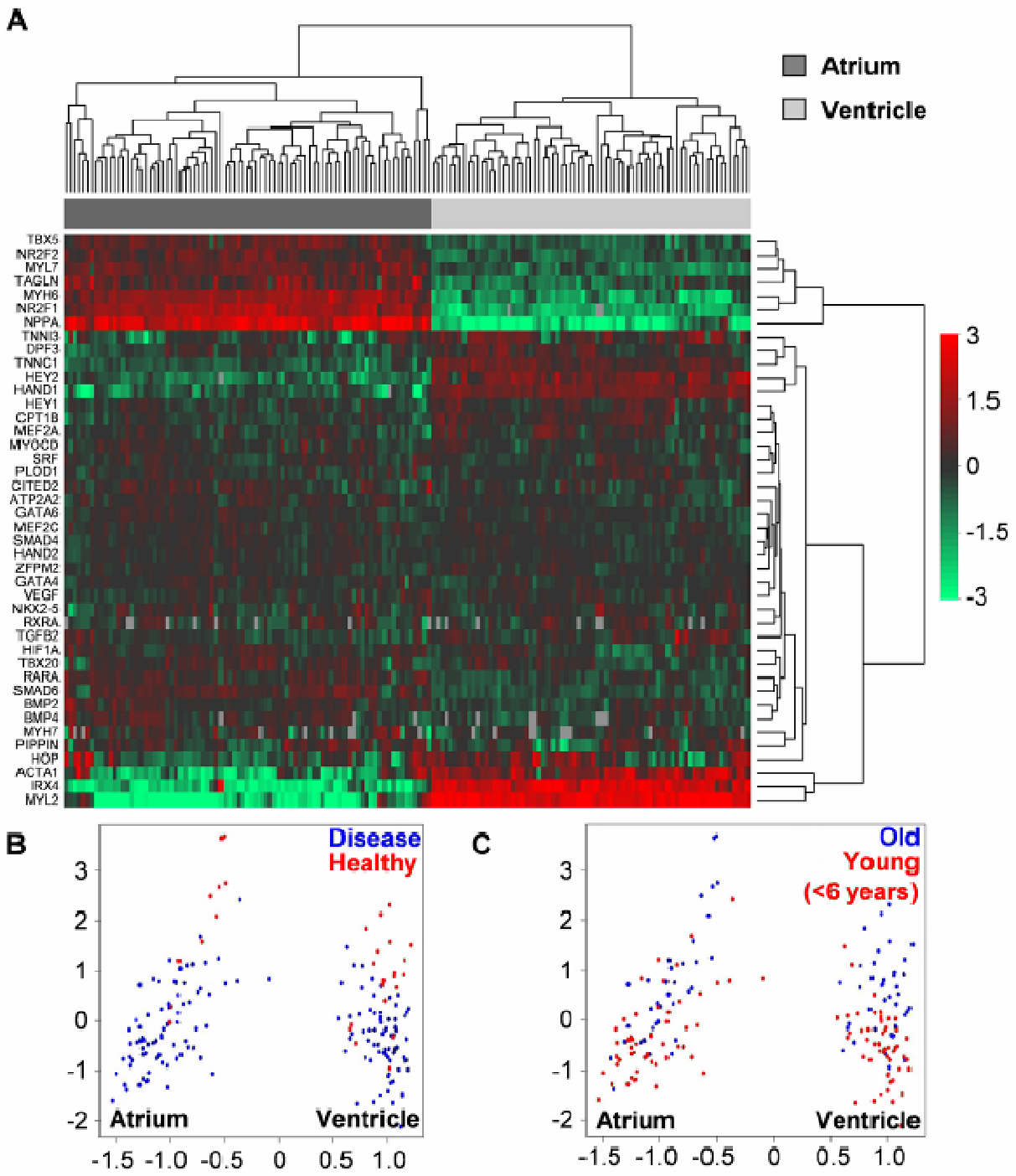

Figure 46: Preliminary gene expression analysis
(A) Hierarchical clustering of gene expression levels measured by quantitative real-time PCR in cardiac samples from patients with different heart malformations. Each column represents a gene and each row a single cardiac sample. Normalized and centered expression levels are color coded in red for up regulated and green for down regulated genes. Missing values are depicted in gray. **(B+C)** Biplot obtained from correspondence analysis. Each dot represents a single patient sample color coded by **(B)** disease state or **(C)** age.

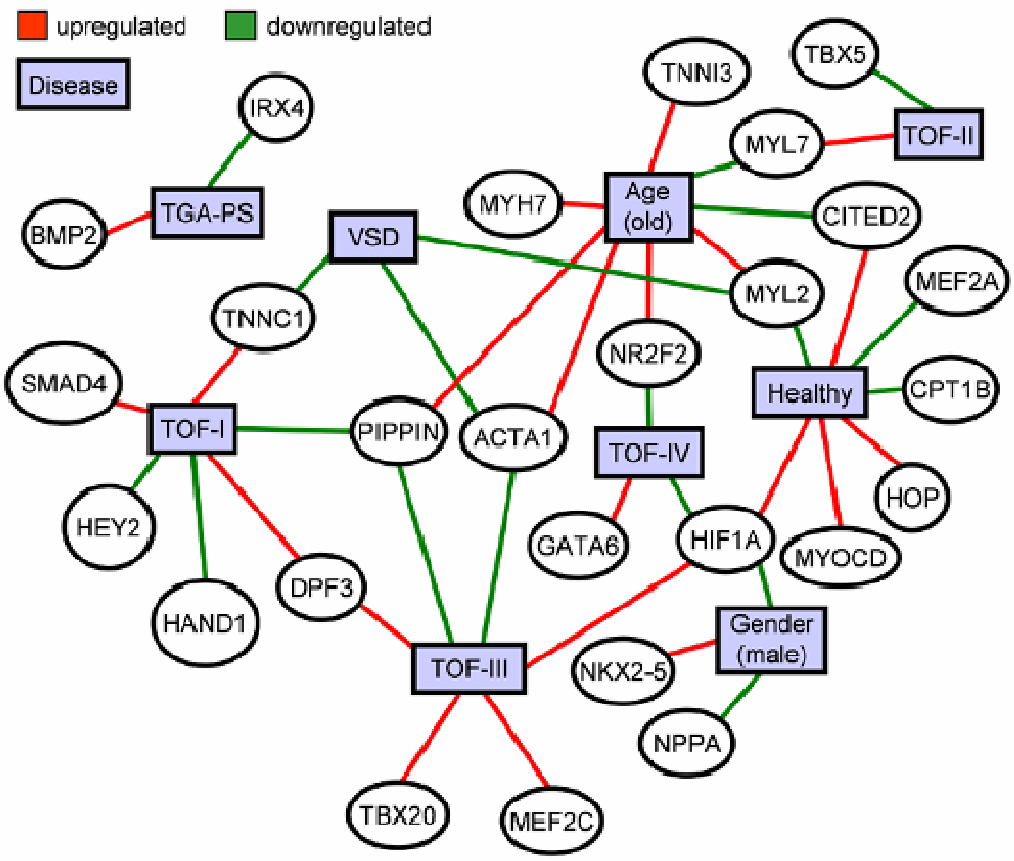

Figure 47: Phenotype to genotype association
Network obtained from linear modeling showing significantly deregulated genes in ventricular samples associated to meta-phenotypes as well as age and gender (marked as blue rectangles). Genes are depicted as circles. Green and red arrows indicate down and up regulated genes, respectively, using a significance level of 0.05.

3.3.4 Linear Model to Detect Disease-Associated Profiles

To identify meta-cluster specific deregulated gene expression, an ANOVA of the form

$$Y = \beta_{meta-phenotype} + \beta_{age} + \beta_{gender} + \varepsilon$$

was estimated, where the observed variable Y represents the gene's expression and $\beta_{meta\text{-}phenotype}$ is the coefficient for each individual patient group sharing the same meta-phenotype. β_{age} was included as the coefficient for the two age categories young and old and β_{gender} to determine gender specific effects. Both have been shown to influence gene expression in the human heart.[78] No intercept term was used because each individual expression vector was centered on zero beforehand. One ANOVA per gene was estimated using least square estimation and the significance of the individual coefficients being different from zero was tested using F statistics. Using a significance level of 0.05 a number of genes were found to have significant coefficients for individual meta-phenotype clusters as well as age and gender. Figure 47 depicts the resulting genotype to phenotype relations as a bipartite graph, where one set of nodes represent the genes (circles) and the other set represents the meta-phenotypes as well as gender and age (rectangles). Green and red edges between a gene and a phenotype indicate significant down and up regulation, respectively. Interestingly, deregulated genes were found for almost all meta-phenotypes, except the cluster *Diverse*, which contains a mixture of different minor

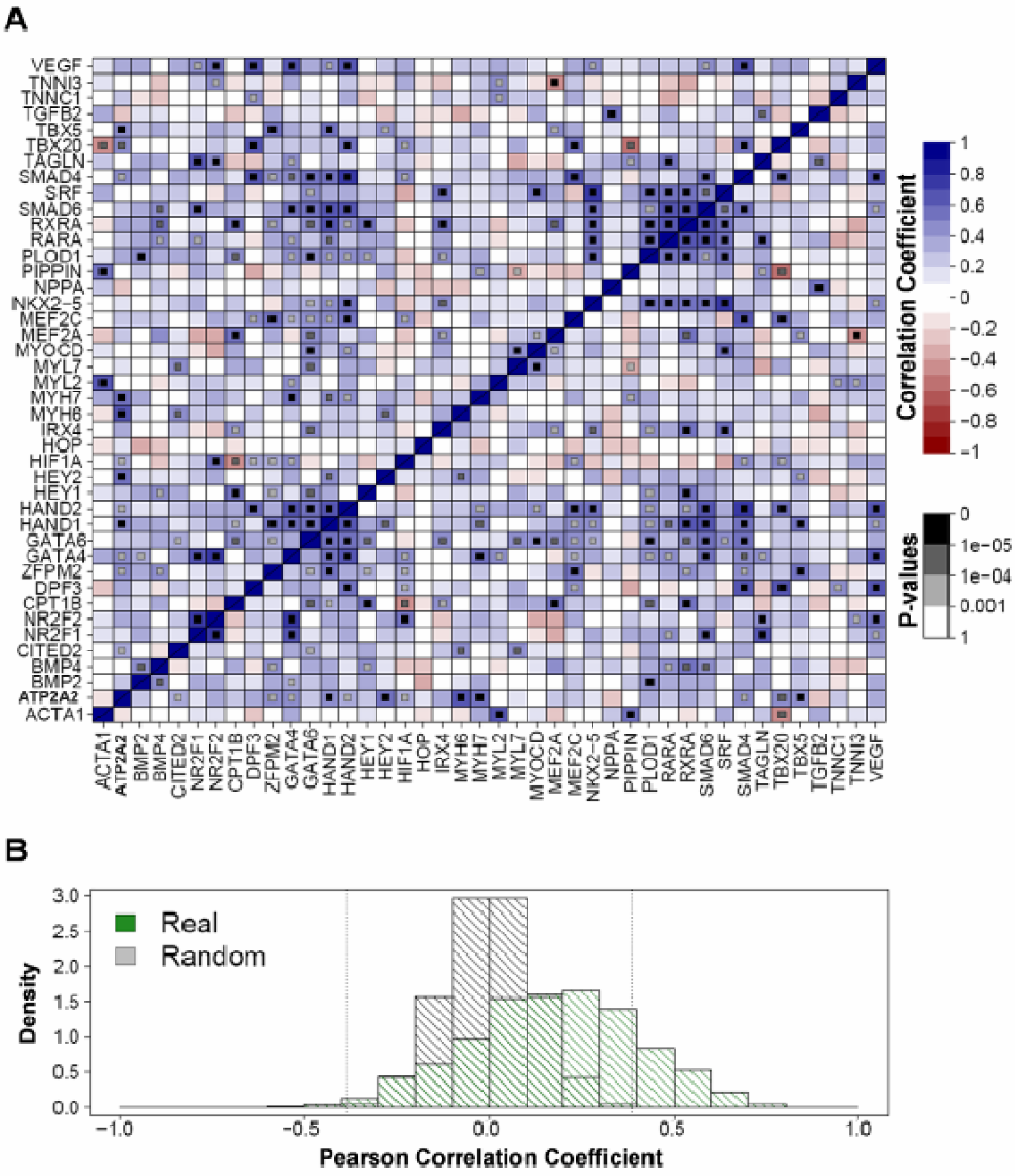

Figure 48: Significant correlation of gene expression
(A) Heatmap of Pearson correlation coefficients and empirical p-values. Computed correlation coefficients are depicted by a blue (positive correlation) to red (negative correlation) color scheme. Small gray boxes indicate empirical p-values. A missing box indicates an empirical p-value > $1x10^{-3}$. **(B)** Histogram of pairwise Pearson correlation coefficients for real and random data.

phenotypes excluding VSD and with regular aortic source from the right ventricle. The other meta-phenotypes, characterized by distinct and moderate to severe abnormalities, have specific molecular portraits, such as *TBX20* and *MEF2C* being up regulated in patients with TOF and main pulmonary artery abnormalities (cluster *TOF-III*), whereas *TBX5* was only down regulated in patients with TOF and bicuspid pulmonary valve (cluster *TOF-II*). While some genes are deregulated in several disease clusters, others appear to be significantly deregulated in all disease samples, indicated by an opposite regulation in the *Healthy* cluster (*e.g. MEF2A* is up regulated in all disease meta-phenotypes).

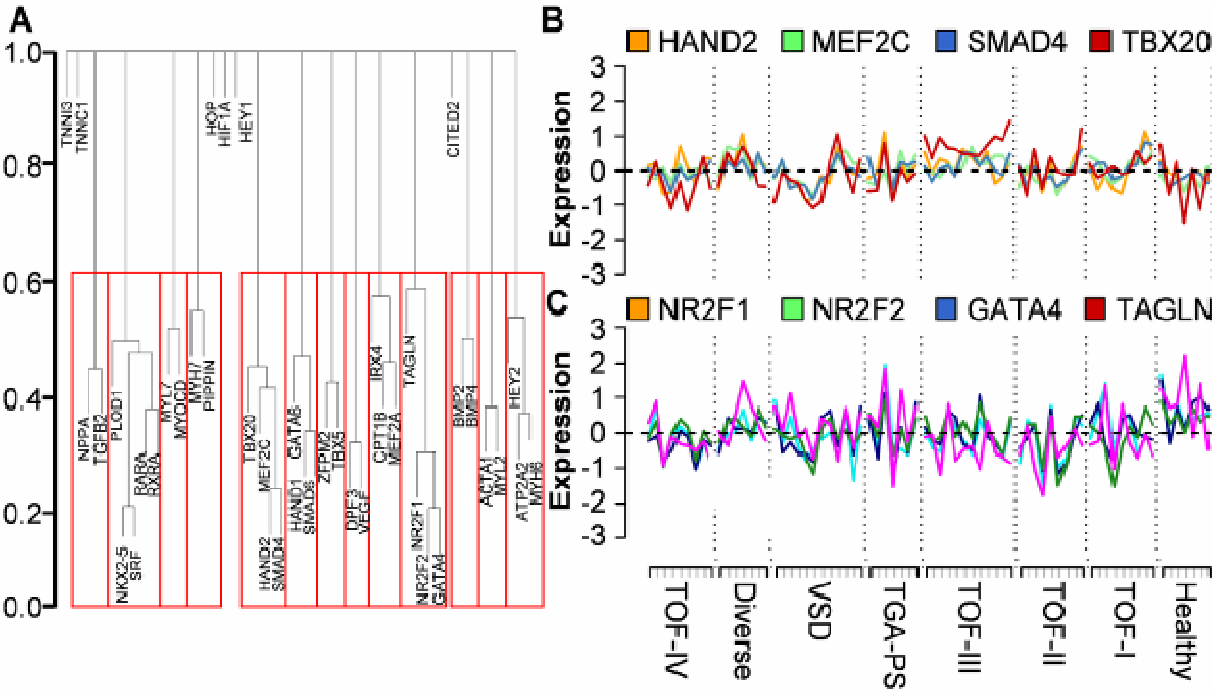

Figure 49: Defined correlated gene groups
(A) Cluster dendrogram showing 13 correlated gene groups. Cluster assignment was derived by cutting the dendrogram at the 0.001 quantile of the random distribution (Figure 48). The *Y*-axis indicates cluster distances. The resulting correlated gene groups are further depicted in Table 16. **(B+C)** Example of two correlated gene groups showing highly correlated patterns of expression in samples of healthy individuals and patients. Centered expression vectors were sorted by defined meta-phenotype.

3.3.5 Defining Correlated Gene Groups

To finally build transcription networks in human, TFBS prediction was applied, focusing on the prediction of measured TFs. However, a simple prediction of *cis*-regulatory elements in the promoters of genes very likely results in many false positive due to the low signal-to-noise ratio, obscuring any subsequent conclusions. A main step to reduce the number of false positive predictions, the predictive power was increased by searching for *cis*-regulatory elements that are shared between promoters of tightly co-expressed and therefore likely co-regulated genes. To find these groups of co-expressed genes, the pairwise Pearson correlation coefficient was computed on the expression data over all samples in the dataset. Like in the analysis of mouse hearts, random experiments were used to evaluate the statistical significance of found correlation coefficients. As a null model, measurements were randomly assigned to samples in the according expression vectors without replacement and Pearson

Cluster	Genes contained in cluster
1	NPPA, TGFB2
2	NKX2.5, PLOD1, RARA, RXRA, SRF
3	GATA4, NR2F1, NR2F2, TAGLN
4	MYH7, PIPPIN
5	HAND2, MEF2C, SMAD4, TBX20
6	GATA6, HAND1, SMAD6
7	TBX5, ZFPM2
8	DPF3, VEGF
9	CPT1B, IRX4, MEF2A
10	MYL7, MYOCD
11	BMP2, BMP4
12	ACTA1, MYL2
13	ATP2A2, HEY2, MYH6

Table 16: Defined correlated gene groups

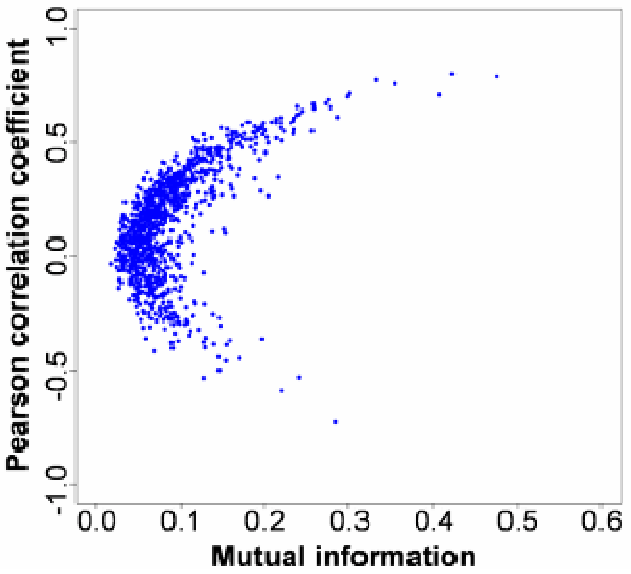

Figure 50: Comparing mutual information to Pearson correlation coefficient
Pearson Correlation coefficients are plotted against mutual information based on kernel density estimation. While there is statistical fluctuation between the two measures, no gene pair is found which has high mutual information but a nearly zero correlation coefficient. Note the different ranges of the two measures.

correlation coefficients were computed on the randomized expression vectors. This process was repeated 100,000 times and the extent of randomized coefficients exceeding the true coefficient was counted. Thereby, an empirical p-value for the measured correlation coefficient of each individual gene pair was derived. A detailed overview of measured correlation coefficients and assigned p-values is shown in Figure 48. To ensure a high level of significance a minimal empirical p-value of 0.001 was required.

Subsequently, complete linkage hierarchical clustering was performed only on significant correlation coefficients, while all non-significant coefficients were set to zero. The 0.001 quantile of the overall distribution from the random experiments was used to cut the clustering tree, thus deriving 19 clusters with significant similarity in gene expression vectors between individual genes (Figure 49 A). Clusters comprising more than one gene were called *correlated gene groups* (Table 16) and two examples are shown in Figure 49 B and C. Centered expression vectors were sorted by the defined meta-phenotype and similar expression patterns of genes are clearly revealed in normal and disease tissue samples. The transcription factors *TBX20* and *MEF2C* displayed correlated expression patterns (Figure 49 B) and strikingly, both are up regulated in patients belonging to the *TOF-III* cluster as analyzed with the linear model.

3.3.6 Comparing Pearson Correlation Coefficient to Mutual Information

For simplicity, the Pearson correlation coefficient was used to define the correlated gene groups. However, different from it's usage for diagnostic reasons as done *e.g.* in the correlation analysis of ChIP-enrichments in mouse hearts, the prediction of *cis*-regulatory elements and the subsequent construction of regulatory networks heavily relied on the validity of the defined correlated gene groups. Therefore, the definition of the right distance measure was a key factor for the analysis. It was shown that the Pearson correlation coefficient can be sensitive to noise effects and outliers[236] and is easily distorted when the expression levels are not uniformly distributed across the expression

Transcription Factor	Associated TRANSFAC Matrices
GATA4	V$GATA4_Q3, V$GATA_Q6
GATA6	V$GATA6_01, V$GATA_Q6
HAND1/HAND2	V$EBOX_Q6_01
HIF1A	V$AHRHIF_Q6, V$HIF1_Q3, V$HIF1_Q5
MEF2A	V$AMEF2_Q6, V$MEF2_02, V$MEF2_Q6_01, V$HMEF1_Q6, V$MMEF2_Q6
MEF2C	V$MEF2_Q6_01
NKX2.5	V$NKX25_01, V$NKX25_Q5
NR2F1/NR2F2	V$COUPTF_Q6, V$COUP_DR1_Q6, V$DR1_Q3
RARA/RXRA	V$DR1_Q3
SMAD4/SMAD6	V$SMAD_Q6_01
TBX5	V$TBX5_01, V$TBX5_02

Table 17: TRANSFAC matrices assigned to TFs present in the dataset.
Matrices removed in the pre- or post-filtering steps have been excluded.

patterns.[124] To determine if any of these effects applies to the definition of correlated gene groups, pairwise mutual information was computed for all gene pairs in the dataset using a density derived definition (section 2.3.3). Comparing the pairwise mutual information for gene expression vectors to the previously computed Pearson correlation coefficients apart from statistical fluctuations an almost one-to-one correspondence was found between these two (Figure 50) pointing to no relevant nonlinear correlations in the data. Therefore, using the Pearson correlation coefficient to define correlated gene groups in this dataset is valid and covers the major fraction of possible dependencies.

3.3.7 Optimized TFBS Prediction using ChIP Data

The last step for in prediction of regulatory networks was the definition of the optimal prediction parameters. The length of promoter sequence as well as the use of conservation information taken for TFBS prediction varies among different studies.[237-239] To make the TFBS prediction as biologically meaningful as possible with regard to these settings, the ChIP-chip data obtained in HL-1 cells for Gata4, Mef2a and Nkx2.5 was used to find optimal prediction parameters. Srf was not used for the optimization as because of the low agreement found between Srf peaks and Srf TFBS predictions using the TRANSFAC[240] matrices.

To find an optimal balance between length of promoter sequence and noise level in the prediction of TFBSs different upstream and downstream distances were used as optimization criteria. Based on transcription start sites in Ensembl (version 48) and in accordance with the array design of the ChIP-chip approach, 10 kb upstream and 3 kb downstream of the 42 selected genes were retrieved from NCBI human assembly (version 36). For the optimization procedure, upstream distances gradually increasing from 200 bp to the full 10 kb and downstream distances from 100 bp to 3 kb were considered, representing the range between using only the core promoter and using the full measured promoter region in intermediate steps.

Beside the amount of promoter sequence, the level of conservation was used as optimization criteria. To assess conservation of promoter sequences, the full mouse human BlastZ alignment from Ensembl (mouse assembly NCBI m37) was used. In addition to the single nucleotide conservation masking provided by the alignment, a 100 bp window was shifted along the promoter regions and windows exceeding a given percentage of conservation remained unmasked. Thresholds ranging from 0% (no

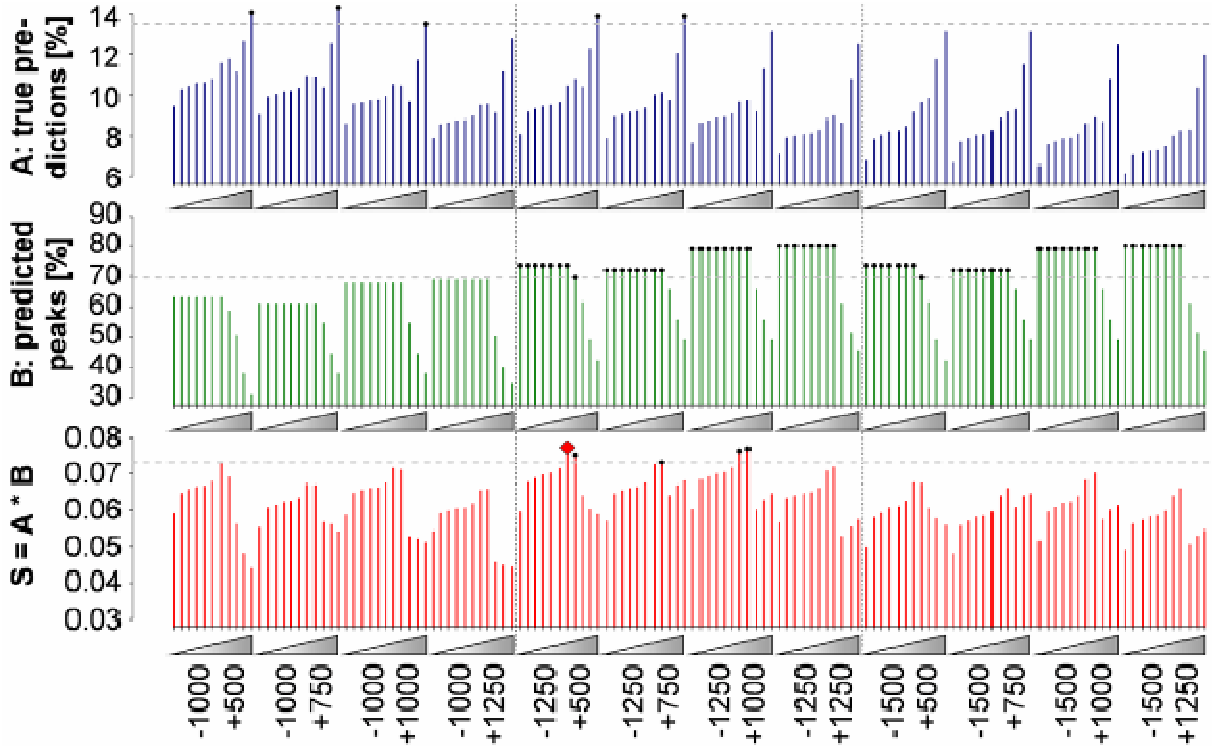

Figure 51: Optimization of TFBS prediction
Results are shown for the TRANSFAC MATCH algorithm and a subset of promoter settings. The upstream (-) and downstream (+) lengths used as promoter are placed below the plot. Triangles indicate the level of conservation ranging from 0% to 100%. Dashed horizontal lines mark best 5 scores, values above this score are highlighted with black dots. The red diamond highlights the best scoring prediction setting.

conservation information) to 100% (single base pair conservation) were evaluated in continuous steps of 10%.

Finally, the prediction of cis-regulatory elements was performed using the TRANSFAC MATCH algorithm (section 2.3.8). 39 matrices representing known binding patterns for 15 of all 22 transcription factors from the heart dataset were retrieved from TRANSFAC (version 11.3) together with their matrix and core similarity scores optimized to reduce both type I and type II error rates. As a pre-filtering step, low quality matrices with a matrix similarity threshold less than 0.8 were excluded, thereby reducing the number of matrices to 27 assigned to 15 TFs. In a post-filtering step, another two matrices showing a very high number of average predictions per promoter were removed. In total this led to 25 matrices associated to 15 TFs which are shown in Table 17. Predictions from matrices belonging to the same TF were combined in order to build the basis for the construction of regulatory networks. To find the optimal TFBS prediction parameters, the scoring function

$$S = \frac{\text{true predictions}}{\text{all prediction}} \times \frac{\text{predicted peaks}}{\text{all peaks}}$$

was used, which was evaluated on each distance and conservation setting. The score S comprises two ratios ranging from 0 to 1 that measure different aspects of the TFBS predictions. The first ratio measures the fraction of true amongst all predictions and the second ratio measures the capability of predicting a ChIP peak. In accordance with the TFBS prediction performed in the analysis of ChIP-chip peaks, a prediction was defined as *true* if it was located not more than 250 bp apart from a

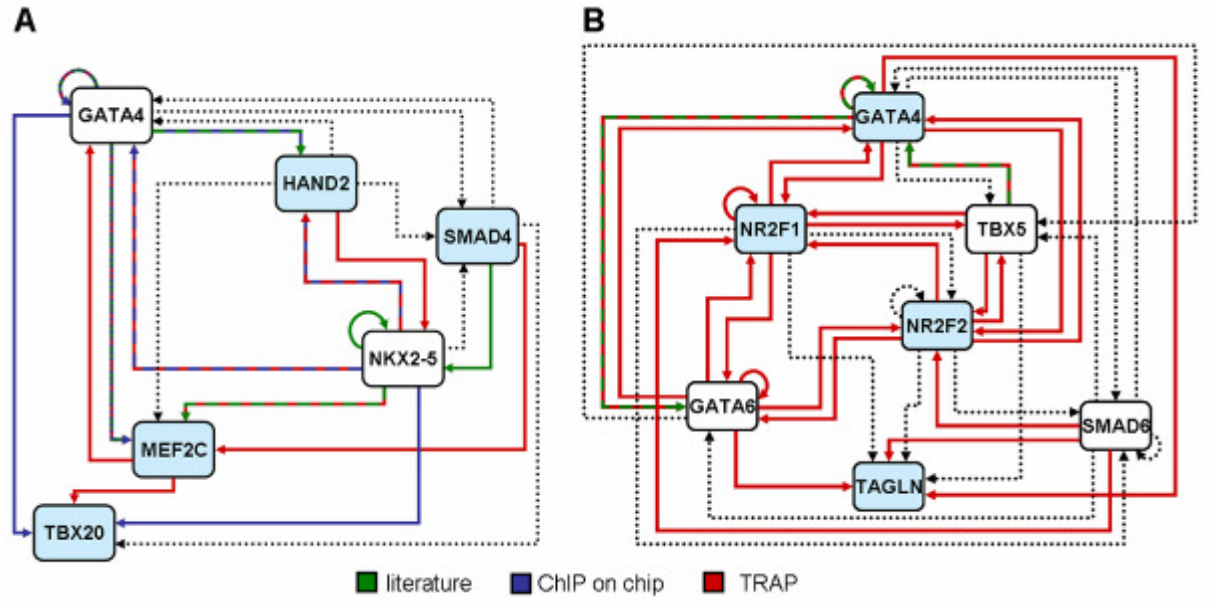

Figure 52: Predicted regulatory networks for two correlated gene groups
Genes composing a group are marked light blue. Confirmation of predicted binding by literature, ChIP-chip and/or TRAP is depicted in colors. Unconfirmed predictions are indicated by dashed lines.

respective ChIP peak center. Furthermore, peaks were marked as predicted if they had at least one true prediction assigned. Predictions as well as peaks were evaluated with respect to the tested promoter settings and peaks lying outside of the evaluated promoter regions were excluded. The optimization process was performed for all three TFs and the average over the three individual scores (one per TF) computed for each setting was reported (Figure 51).

Studying the individual scoring results, it could be observed that the fraction of true predictions decreased with the length of the sequence used, which is in line with the proposed increase in noise level. However, binding sites identified by ChIP can be observed at any distance from the transcription start site. Further, while the fraction of true predictions could be enhanced by using more stringent conservation settings, the amount of TF ChIP peaks predicted heavily dropped at higher conservation levels like found in the analysis of TFBS in ChIP-chip peaks. This finding is supported by the observation that actual binding sites of TFs might be slightly modified during evolution for example to enable adaptation of TF binding.[241,242] Using the proposed scoring function which incorporates both measures, prediction settings of 1,250 bp upstream and 500 bp downstream together with a conservation level of 60% were found to be optimal for the analyzed TFs when using the TRANSFAC MATCH algorithm for TFBS prediction.

3.3.8 Predicting Cardiac Regulatory Networks

Finally, regulatory networks were constructed based on the identified correlated gene groups and the optimized transcription factor binding sites prediction representing the underlying regulatory dependencies. This resulted in several small subnetworks, which could subsequently be analyzed in more detail. For verification the constructed networks were compared with binding data from ChIP-chip and literature data. Figure 52 displays two graphs representing predicted regulatory subnetworks for the correlated gene groups comprising *HAND2*, *MEF2C*, *SMAD4*, *TBX20* (Figure 52 A) and *GATA4*, *NR2F1*, *NR2F2*, *TAGLN* (Figure 52 B).

For the first group, GATA4 and NKX2.5, which are known to interact with each other, were directly predicted to bind all four promoter regions and all except the two bindings to SMAD4 have been proposed in literature (Nkx2.5→*Mef2c*[243,244]), found in the ChIP-chip data (Nkx2.5→*Hand2/Tbx20*, Gata4→*Tbx20*) or both (Gata4→*Hand2*[245]/*Mef2c*[246]). Interestingly, both *TBX20* and *MEF2C* are specifically up regulated in patients belonging to the *TOF-II* cluster and the prediction approach sheds light on potential upstream regulators.

Concerning the *GATA4, NR2F1, NR2F2* and *TAGLN* correlated gene group, several TFs were found that had predicted binding sites in all promoter regions of the four genes. Among them are TBX5, GATA6 as well as GATA4 and the two NR2F factors. Identification of the latter three is quite remarkable as all three TFs present in this correlated gene group show inter-regulatory interactions that could explain the observed correlation. Some connections have already been described in literature (Gata6→*Gata4*[247]/*Tagln*[248], Gata4→*Gata4*[247]) but no binding was found in our ChIP-chip data. However, it should be kept in mind that the ChIP experiments were performed using mouse cardiomyocytes (HL-1 cells), whereas the predictions are based on transcription patterns from human patient material.

Finally, in order to substantiate the predicted TF regulations, the transcription factor affinity prediction (TRAP; section 2.3.8) algorithm developed by Roider *et al.*[183] was incorporated. In contrast to TFBS prediction of the MATCH algorithm, the provided affinity measure is continuous and allows an easy ranking of promoter regions with the highest affinity for each TF matrix. To compare the results gathered from the optimized TFBS prediction with TRAP, the 10 promoter regions with the highest affinities for each TF were marked as potentially regulated by this TF.

Applying TRAP to the correlated gene group comprising *HAND2*, *MEF2C*, *SMAD4* and *TBX20*, no TF was found which had high affinity for all four gene promoter regions. Remarkably, SMAD4 was no part of any of the top 10 affinity promoter regions of any TF analyzed, although the SMAD4 promoter was predicted to be bound by a large fraction of TFs (Figure 52 A). Regarding the results of the TFBS prediction, NKX2.5 was assigned by TRAP to two of the remaining three genes, namely *MEF2C* and *HAND2* (confirmed by literature and ChIP-chip, respectively), but did not show high affinity to *TBX20*. However, binding of Nkx2.5 to *Tbx20* was observed in ChIP-chip. Therefore NKX2.5 is very likely a crucial factor for the stated correlation.

In case of the *GATA4, NR2F1, NR2F2* and *TAGLN* correlated gene group, both GATA4 and GATA6 appear to have all four gene promoter regions in their top 10 affinity tables. This underlines the results of the TFBS prediction in which they also showed binding to all group members. Furthermore, it highlights GATA proteins as potential auto-regulatory key factors in the given subnetwork. In addition, SMAD6 showed high affinity to three of the four correlated genes, namely *NR2F1, NR2F2* and *TAGLN* and was predicted to be bound by GATA4 itself, which implies a functional role further downstream in the regulatory cascade.

3.4 Implementing the CArdiovascular Regulatory INteraction Database

The main conclusion that can be drawn from this study is that the construction of transcriptional regulatory networks can only be successful if it is based on a range of complementing experiment and integrates a multitude of different data sources. The main problem, however, is to gather this data and to combine available annotations in a biological meaningful manner. Therefore, as a final step in this

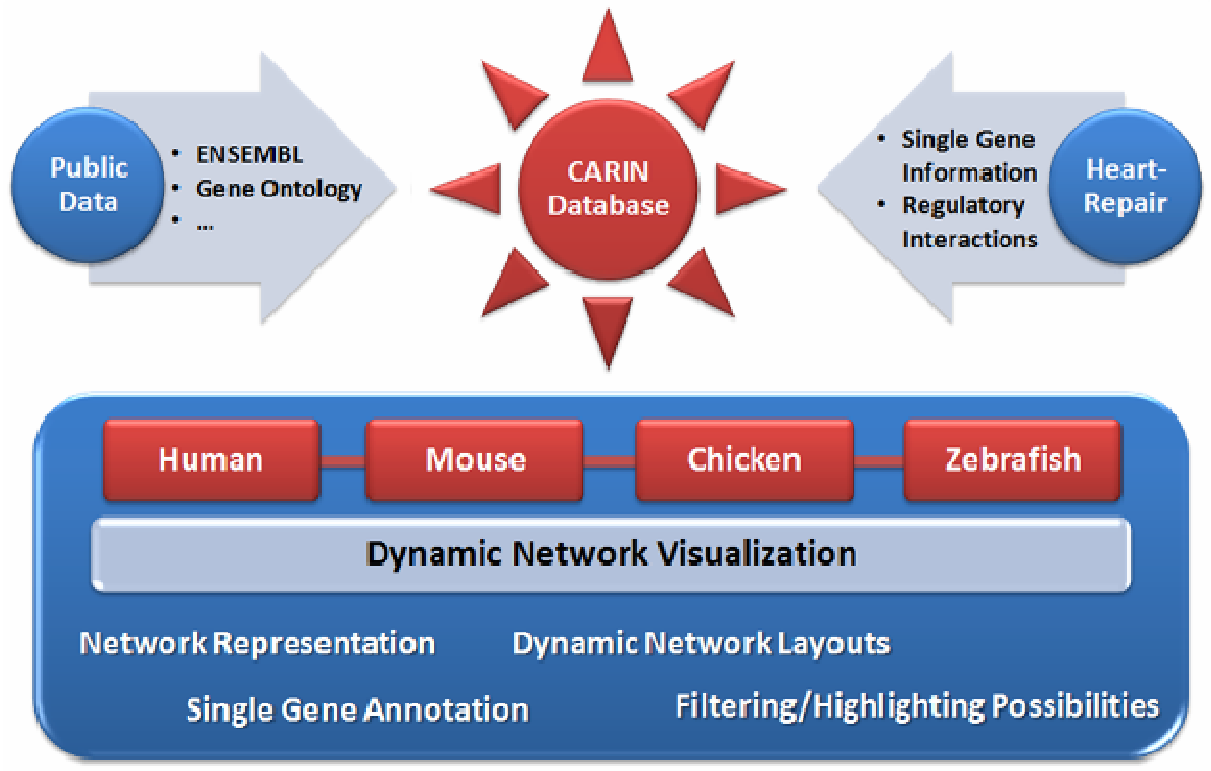

Figure 53: Scheme showing the data architecture of the CARIN database and its user interface

study, CARIN, the CArdiovascular Regulatory INteraction database, was implemented. CARIN is a program that integrates a database of up-to-date experimental and computationally derived annotations relevant for cardiac and muscle genes with a sophisticated user interface that provides an easy and comfortable data overview and detailed information for individual genes at the same time. It was developed as the dissemination database for the HeartRepair EU project and incorporates data of a number of HeartRepair research groups and interlinks these to publicly available databases incorporating Ensembl and the Gene Ontology project.

3.4.1 General Purpose

The previous analyses have revealed that to understand molecular and developmental pathways in eukaryotic cells, transcription factors must be viewed within their regulatory context. This includes the interplay between transcription factors and co-regulatory elements but also epigenetic factors such as histone modifications. Further on, the mere binding of a transcription factor to a gene promoter does not imply active control of transcription and therefore knockdown experiments are necessary to identify a direct regulatory impact. So far large public datasets comprising all these different aspects of transcriptional regulation over whole genomes from higher organisms are still rare.

In an attempt to make existing data available and interpretable for future studies CARIN, the CArdiovascular Regulatory INteraction database, integrates data obtained by members of the EU-funded project HeartRepair (LSHM-CT-2005-018630) into a common framework. Studied gene expression profiles and transcription factor binding events from human, mouse, chicken and zebrafish are implemented and provided together with gene associations based on literature mining an information resources for cardiovascular regulatory networks. CARIN provides information on a gene-wise level using the Ensembl[106] database as reference. It enables the querying for genes of interest and

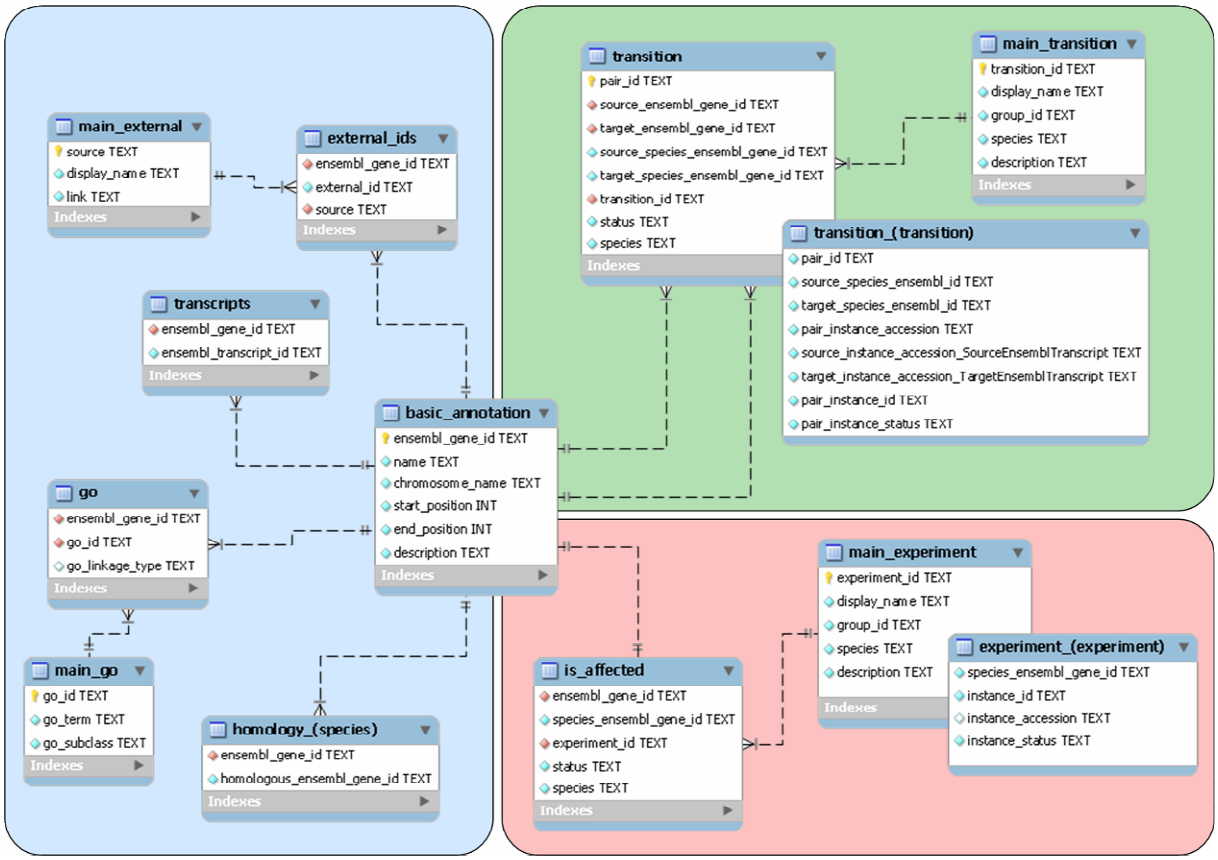

Figure 54: Relational scheme of the CARIN database
Outline of the relational database scheme. Tables are grouped according to their data sources: Ensembl (blue), single gene experiments (red) and relation experiments (green). Each table with brackets in its name stands for a number of tables, one per experiment/species. The relational scheme was drawn using the MySQL Workbench program version 5.2.

places individual datasets into a broader context. Detailed experimental annotations are provided by the contributing groups to allow a well-grounded interpretation of the data. In addition, annotations from or links to publicly available databases like Ensembl, the GeneOntology[177] consortium or the UCSC Genome Browser[185] are given. Genes are linked either by studied transcription factors binding events or based on information generated by the literature and data mining tools Anni[249] and STRING[250]. A high value was set on the visualization aspect. A dynamic network representation has been implemented to provide an easy and comfortable data overview and detailed information for individual genes at the same time. Further, the system incorporates several features for highlighting and filtering genes and interactions, which provides the possibility to focus on specific regulatory pathways or fields of interest.

3.4.2 Data Architecture

CARIN is build of two layers, one comprising the database, which itself is split into several sub-databases, and the user interface which processes the data from the individual database into a network representation and implements features like the possibility to filter or highlight genes of interest and to present annotation on a gene-wise level (Figure 53).

The Database

CARIN was developed to incorporate gene annotation and regulation data from several data sources and was implemented as a relational database (section 2.3.9) with genes as the main entities. To allow the integration of different experimental and computational approaches, Ensembl gene IDs were taken as a reference to represent individual gene entities. In this way, identifiers of any data source that is going to be integrated into CARIN must be mapped to the appropriate Ensembl species database. This approach was chosen because Ensembl supplies a reasonably stable and publicly free platform and many other databases and companies therefore provide the appropriate Ensembl gene IDs for their own identifiers, making a complicated mapping unnecessary for the user. In addition, Ensembl itself provides a large number of pre-integrated data sources, like GO annotations or mappings to other free databases, and offers a free API-based web-access to all their data, which allows an automatic update using appropriate scripts. Further, Ensembl provides a homology mapping between genes from different species, which is crucial for the integration of data gathered from different organisms.

Figure 54 shows the relational database scheme, which is divided into two parts with the main table "*basic_annotation"* as the center. In the left (blue) part, tables comprising gene annotations directly retrieved from Ensembl are shown. These comprise the GO term associations and external identifiers. In the right (green/red) parts, tables comprising annotations retrieved from experiments conducted from any of the HeartRepair project members as well as the Anni and STRING tools are shown. The experiments themselves divide into two further groups: One group (green) represents experiments that provide annotations for a single gene, *e.g.* transcriptional profiling of a number of genes under certain conditions. The other group (red) represents relations between genes as gathered from ChIP experiments or co-occurrence in literature. To make the submission of experiments as easy as possible for the HeartRepair contributors a single information sheet which is highly similar to submission sheets of the GEO[251] and ArrayExpress[252] databases was developed to submit experiments of both kinds. The information that had to be contributed with each submission is given in Table 18. While CARIN in general stores the annotation for all genes from a submitted experiment in its database, only those genes that are significant in the submitted experiment according to the contributor are represented by the user interface. Due to the fact that the decision of significance is made by the contributor and is not decided according to a fixed significance level or based on a fixed statistical model, it is possible to integrate data from many different experiments that might be performed under many different conditions by making use of the contributor's knowledge of the individual experiment.

CARIN was developed to integrate experiments performed in several species, thus a concept to exchange data between the different species was needed. In general, two oppositional ways of data integration from several species are conceivable: the first approach is to store all annotations using a single pre-defined reference species and all annotations must be mapped to gene identifiers of that single species. The second approach is to define overall gene entities, which are species-independent, and map annotations from any species to these comprehensive identifiers. While the first approach is simple and straight-forward, it provides the possibility to loose a high amount of data, if the experiment species and the reference species are very dissimilar. It further puts high influence on the choice of the reference species as different organisms are more and others are less similar to each other regarding their genomic content. The second approach copes with this problem by using identifiers that are species independent. However, the definition of reasonable entities is a challenging task as a gene in one organism might *e.g.* been duplicated during evolution, which could lead to two paralogs with altered functions in another organism.

Experiment Info	
Experiment name	Descriptive name for the experiment
Contributor name + email	Contact info of the submitter
Contributor group	Name of the research group
HeartRepair group	HeartRepair contract number
Description	Description of the used experimental model
Comparison listed^{TP}	Conditions that were compared in the experiment
Reference^{TP}	The reference condition
Tissue type	Tissue that was used for the experiment
Strain^{N}	Strain of the organism used in the experiment
Species	Organism used in the experiment
Treatment^{N}	Treatment of the animals used in the experiment
Experimental design	Short description of the experimental design
Technological design	Short description of the technological design
Array platform^{N}	Name of the chip manufacturer
Chip id^{N}	Array version used
Number of probes^{N}	Total number of probes present on the array
Score details	Scale of the score in the data table
P-value details	Statistical model used to calculate the p-values
Submission date	Submission date
PubMed id^{N}	Publications associated to the data
Array-express/GEO identifier^{N}	Public database identifier for the data

Data Matrix	
ID	Unique probe identifier
EntrezGene ID	EntrezGene identifier for the measured transcript
EnsemblTranscript ID	Ensembl transcript identifier for the measured transcript
P-value	P-value for the differences between the conditions tested
Score	Fold change difference between the conditions tested
Significant	Indication if the transcript is considered to be significantly different between the conditions tested (0/1-coding)

Table 18: Submission fields for the CARIN database
Experiment submission to CARIN are done using a tabular sheet that is structured into two parts. The **experiment info** contains information for the performed experiment, the contributor and the methods used for the analysis. The **data matrix** contains one row per measured probe and includes resulting scores, p-values and the indication of significance. While the submission is based on transcripts, these will later be mapped to genes in the CARIN database. N = can be empty; TP = only for transcriptional profiling

To benefit from the simplicity of the first model and keep as much data as possible at the same time, CARIN implements an approach that lies in between these two models. Instead of using a single reference species for the whole database, a single database is build for every individual species with submitted experimental annotations. Using the homology mapping from Ensembl, experiments performed in other species are then mapped to each reference species in turn, allowing only one-to-one homology mappings to ensure the integrity of the database and transferability of knowledge between the two species. It is then left to the user to decide which reference species is used in the CARIN session in order to maximize the output in accordance to the needs of the analysis.

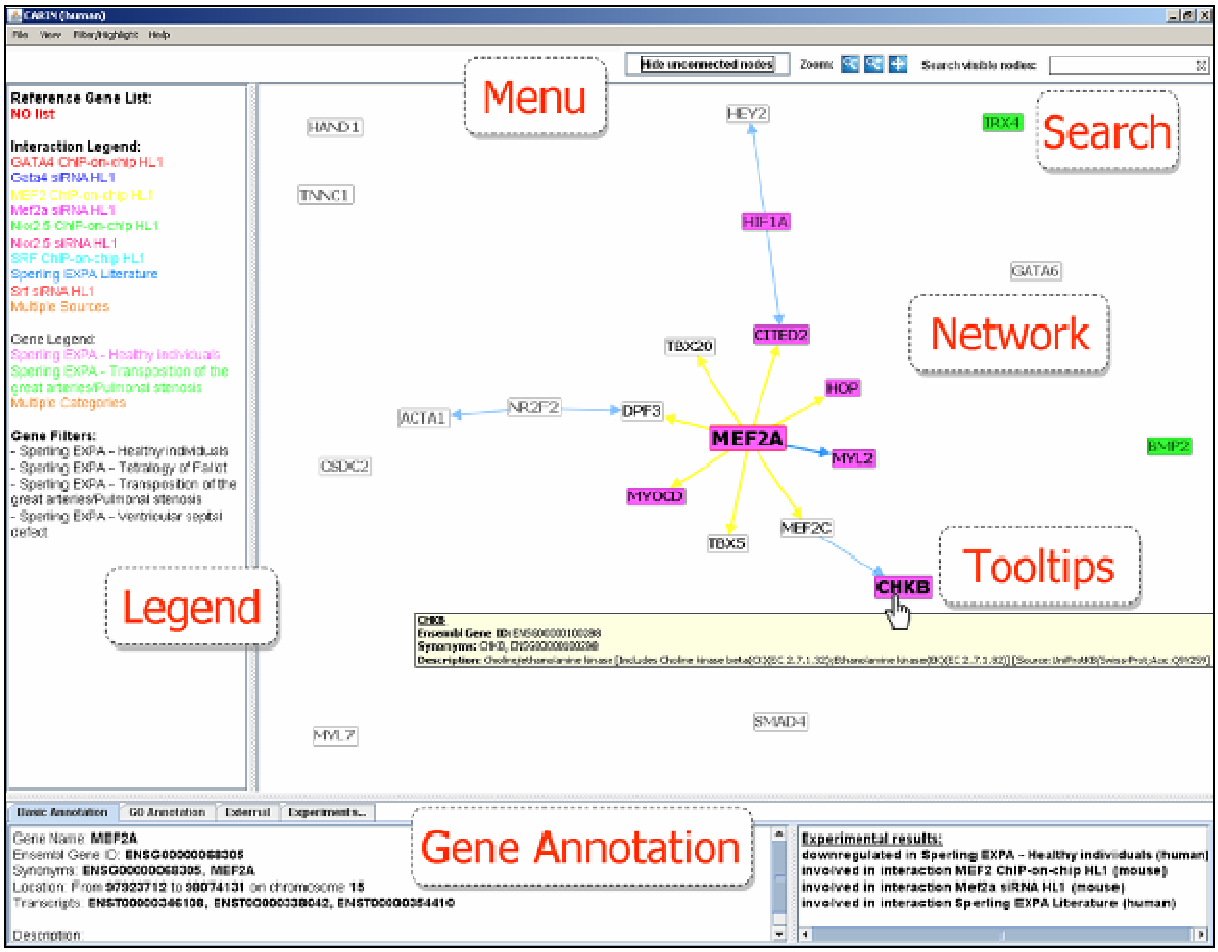

Figure 55: The CARIN user interface

User Interface

The user interface is the main part of CARIN. It is structured into two subparts: the network part and the gene annotation part (Figure 55).

The network part represents genes and annotated interactions in form of a network. Nodes represent individual genes that appeared to be significant in any of the experiments included in CARIN. Edges between these genes reflect regulatory interactions (*e.g.* binding in ChIP or strong association according to literature mining). The used network layout algorithm is an implementation of the radial tree layout from Yee *et al.*[253] and arranges genes around the center node on concentric circles using the path length to determine the distance to the center. In this way, genes that are related to the center will be displayed in its proximity. Genes not connected to the center node are randomly placed around the network and can be hidden to focus only on genes functionally coupled to the center node. In general, information is color-coded, with a specific color for every experiment in CARIN. To change the appearance of the network, genes can be dragged and the network can be zoomed. Additionally, double clicking a gene marks it as the new center node of the network, computes a new network layout and changes the information in the gene annotation part. A prefix based search which highlights all matching nodes with a specific color allows the user to identify specific genes of interest. Tooltips provide short information for genes and connecting edges.

The gene annotation part comprises all annotations for the current center gene divided into several panels. It displays the gene name, Ensembl ID and other Ensembl related information together with all

experiments where the gene is marked as significant. In addition, it shows GO terms associated to this gene, external identifiers and links to their appropriate online databases. To allow an easy interpretation of the results, annotations for all experiments in which the gene was marked as significant are listed.

To focus on specific sub-networks or experiments, the user interface contains two filtering systems. The experiment-based filtering allows the definition of gene and interaction filters based on any of the submitted experiments. With an active filter only genes or interactions significant in the selected experiment are displayed in the network. If several experiments have been selected for filtering, only genes or interactions significant in at least one of these are displayed. By default, no interaction filters are defined, thereby showing all interactions. Another filtering option is the definition of a reference list that contains Ensembl gene IDs of the reference species. If such a reference list is defined, only genes part of that list will be displayed. The experiment-based and reference list filtering can be combined to define filtering rules of any complexity. In addition to the filtering system, the user interface also provides a highlighting system. Similar to the experiment-based filtering, each node and each interaction can be colored according to its significance in selected experiments. If a gene or interaction is significant in two or more experiments a default color is used. A legend next to the network indicates each used filtering and highlighting (left side of Figure 55).

To export data from the user interface, several export formats were implemented. First, it is possible to store the currently visible network as an image file. All the networks shown here have been produced using this option. Second, a text file containing a list of Ensembl gene IDs can be exported together with their gene names for all genes currently visible in the user interface or to exchange data between CARIN and other applications in a readable format. Any highlighting will be included in the file by additional columns, one for each experiment, and 0/1-coding for the significance. These text files can subsequently be used as reference lists for the user interface. Finally, it is possible to save the currently visible graph using the GraphML[254] XML language. Edge source definition and any selected gene highlighting will be included in the file using individual keys and integer values depicting the coding. The application of both filtering systems as well as the highlighting and export properties will be demonstrated in the example session described below.

3.4.3 Implementation

CARIN was implemented as a standalone application and was distributed together with the appropriate libraries and reference species databases to all contributors from the HeartRepair project. It was implemented in two steps: the first step was creating the database and entering the public and experimental annotations. The second step was the implementation of the user interface and its connection to the databases.

Relational Database Setup

Newest gene annotations from Ensembl were gathered using scripts that directly access the aforementioned Ensembl web-API. In addition, annotations of submitted experiments were integrated into the CARIN database using routines that directly access the submitted sheets. Both scripts were implemented in R.[109] The communication between the scripts and the SQLite databases was established using the RSQLite package from Bioconductor.[110] GraphML XML files were subsequently

produced from the resulting databases comprising an abstract representation of the resulting regulatory network and annotations.

User Interface Implementation

The implementation of the CARIN user interface was done in Java using the *prefuse*[255] visualization framework which is also implemented in Java. Prefuse supports a number of features for data modeling, visualization and user interaction and provides optimized data structures for tables, graphs and trees together with a set of layout and visual encoding techniques and support for animation. Further, through its modular structure it is easily expandable and adjustable to different visualizations. CARIN was implemented using a mixture of classes directly extended from prefuse and completely self-developed classes together with a *Swing*-based graphical user interface. The GraphML XML files produced from the individual databases are used to load the network structure into the application, while single gene annotations are directly retrieved from the SQLite databases of the reference species.

3.4.4 Currently Stored Data

Currently, CARIN contains 66 experiments performed in four different species, namely human, mouse, chicken and zebrafish. All annotations were based on Ensembl version 57. From the 66 contributed experiments 51 provide annotations for individual genes, like expression restricted to certain tissues or regions of the heart, expression changes according to treatment with particular substances or the associations to specific congenital heart disease or patient groups retrieved in section 3.3. Another 15 experiments provide annotations on interactions, like direct binding of a transcription factor to the promoter of a target gene as derived by chromatin immunoprecipitation or regulatory dependencies as derived from differential expression analysis in knockout or knockdown experiments including the experiments analyzed in section 3.1.

In addition, CARIN contains interactions based on two bioinformatic approaches. The first is the STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) database, which was developed by Snel *et al.*[250] in 2000 and has been continuously evolved since then.[256-259] STRING is a database dedicated to known and predicted protein-protein interactions, including both physical and functional interactions. It weights and integrates information from numerous sources, including experimental repositories, computational prediction methods and public text collections. STRING integrates interaction data for a large number of organisms and transfers information between these organisms where applicable.[259] The advantage of integrating gene and protein interactions from STRING into CARIN was its broader scope incorporating also non-cardiac interactions, which might facilitate the uncovering of regulatory associations hitherto unknown in the cardiac field. For CARIN, only interactions derived from genomic co-occurrence, co-expression, experimental evidence, literature mining and curated databases were integrated that had an association score of more than 0.7, which is referred to be a *high confidence* score threshold by the developers of STRING. Further, CARIN integrates interactions from Anni[249,260] a literature mining tool that uses concept profiles to find associations between genes. A concept profile is a list of concepts like '*prostate cancer*' which each have a specific weight that reflects its level of association to the gene. Given a predefined set of these concepts, the weights are computed from Medline abstracts either using automatic concept recognition

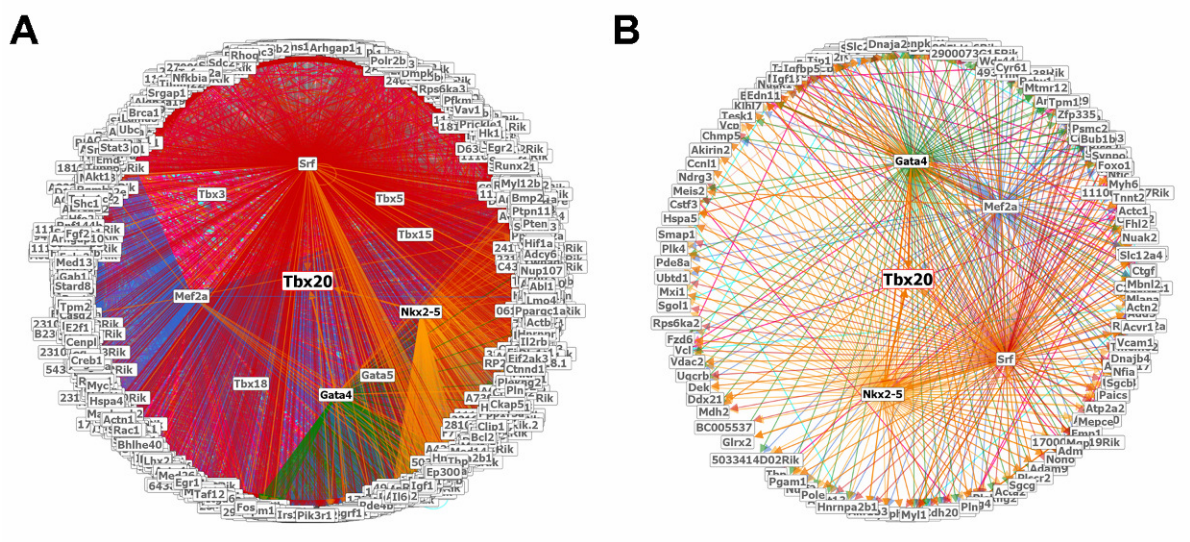

Figure 56: CARIN Example Session (1)
(A) Full network of all genes with at least one TFBS for Gata4, Mef2a, Nkx2.5 or Srf in ChIP-chip. Tbx20 was selected as the center gene. **(B)** Reduced version of the same network with only genes that are additionally differentially expressed in at least one of the respective knockdowns.

or manual curation.[249] To find associations between gene pairs an association score based on the cosine angle between their individual concept profiles is computed. Anni has successfully been used to infer functional associations between genes[261] and between genes and GO terms.[262] For CARIN, only associations with an association score higher than 0.05 were incorporated. The score threshold was suggested by one of the Anni developers and should provide a good balance between very obvious and too far-fetched associations.

3.4.5 Example Session

In the following, a CARIN example session is presented to demonstrate its features, the integration of several experimental datasets and its benefit for the user. The reference species selected for this example session was mouse. As a starting point, a filter based on all included ChIP-chip experiments analyzed in this study was created. Thus, Figure 56 A shows all genes that are found to have a TF peak in their promoter for any of the four analyzed TFs Gata4, Mef2a, Nkx2.5 and Srf. Without an edge filter any connecting edge is displayed, leading to a very large network. The T-box protein Tbx20, already analyzed in the previous sections, was selected as the center gene and is bound by all four factors in ChIP. In addition to the four binding TFs, Figure 56 A shows a number of genes in the proximity of Tbx20. These include other T-box proteins which are connected though Anni textmining associations (Tbx3, Tbx5, Tbx15) or through differential expression in a respective knockdown (Tbx18). Further, Gata5, a co-factor of Tbx20, is associated via STRING textmining associations.

To narrow the number of interesting genes, the gene list of Figure 56 A was exported into a text file using the export function of CARIN. This list is subsequently used to set a reference list filter which will display only genes bound by at least one of the four TFs in the network. A filter for genes that are differentially expressed in any of the four respective siRNA knockdown experiments generates the network in Figure 56 B. All of the genes proximal to Tbx20 in Figure 56 A are lost except the four analyzed TFs, as none of them appears to be differentially expressed as well as bound by any of the four factors at the same time.

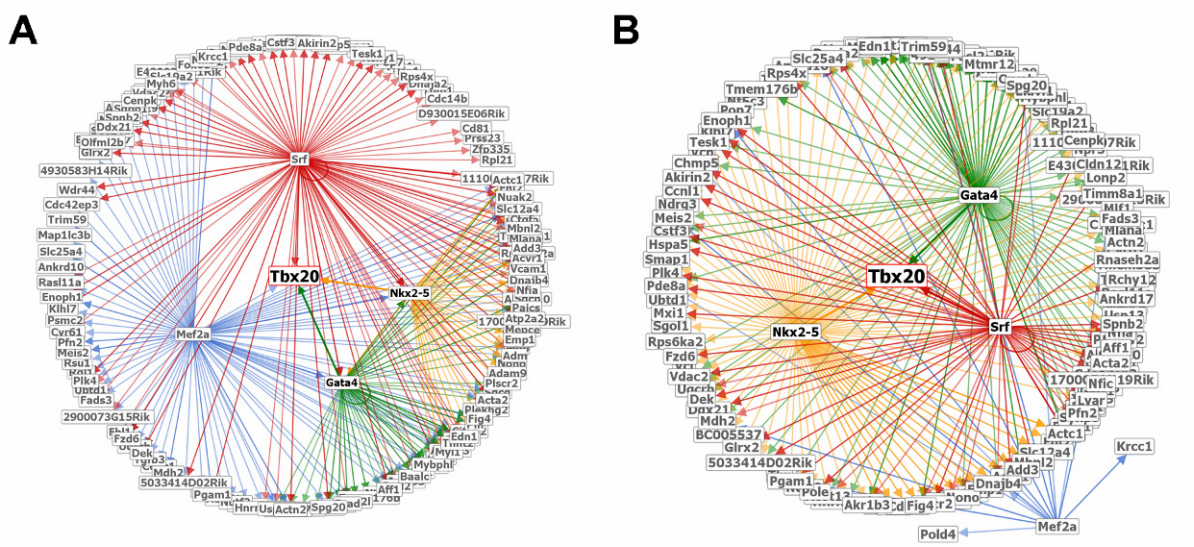

Figure 57: CARIN Example Session (2)
(A) Network from Figure 56 B showing only edges from ChIP-chip experiments. **(B)** Network from Figure 56 B showing only edges representing differential expression in siRNA knockdown.

As the number of edges is still very large, Figure 57 A shows a subnetwork with only those edges that represent binding in ChIP-chip. Every gene displays an incoming edge because the ChIP-chip experiments were used as the reference list. In line with the results from section 3.1, Nkx2.5 and Gata4 share most of their targets while Srf appears to be more detached from the other three TFs. Figure 57 B focuses on regulating interactions instead of promoter binding by showing only edges that represent differential expression in the respective knockdown. Now Mef2a is decoupled from Tbx20, because Tbx20 was found to be differentially expressed in the knockdown of Gata4, Nkx2.5 and Srf, but not in the knockdown of Mef2a. In this way, Figure 57 represents two different networks that combine both the ChIP experiments and the siRNA knockdown experiments analyzed in section 3.1.

Instead of using edges from either ChIP-chip or siRNA experiments, Figure 58 A shows the network of genes that are connected to Tbx20 by STRING experiment and database associations (pink edges), STRING textmining associations (light blue edges) or both (orange edges) and are in addition differentially expressed in any knockdown (filter option) and bound by any of the TFs in ChIP-chip (reference gene list option). To focus only on genes coupled to Tbx20, all genes that had no connection to Tbx20 were hidden from the network using the appropriate functionality of the user interface. The resulting network comprises a much smaller set of 28 genes which could successively be analyzed in more detail. The highlighting represents differential expression in the siRNA knockdown experiment of Gata4 (green), Mef2a (blue), Nkx2.5 (yellow) and Srf (red) or in multiple of these (orange). Interestingly, most of the genes appeared to be differentially expressed in the knockdown of Gata4, which is known to physically interact with Tbx20.[263] To further investigate the regulation by Gata4, Figure 58 B shows the same network with a different highlighting, now representing binding by Gata4 in ChIP-chip (light green), differential expression in the Gata4 siRNA knockdown (green) or both (orange). Many of the genes including Tbx20 show both properties, implying a functional network regulated by Gata4, which could be analyzed in subsequent experimental studies. However, it has to be kept in mind that this small network is the result of reducing the network of all genes that have a STRING association path to Tbx20 to those that are bound by at least one TF in ChIP-chip and are differentially expressed in at least one siRNA knockdown. Removing one of this two filter options would result in much larger networks like the one

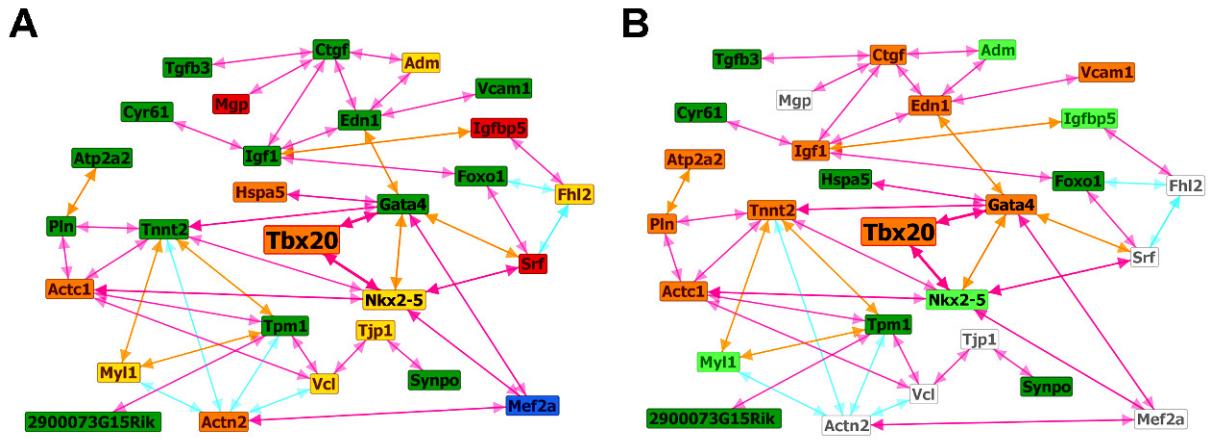

Figure 58: CARIN Example Session (3)
(A) Network from Figure 57 B showing only STRING experiment and database associations (light blue), STRING textmining associations (pink) or both (orange). All genes not connected to Tbx20 have been removed from the network. The used highlighting represents differential expression in the knockdown of Gata4 (green), Mef2a (blue), Nkx2.5 (yellow), Srf (red) or in multiple knockdowns (orange). **(B)** The same network now with a different highlighting indicating binding by Gata4 in ChIP-chip (light green) or differential expression in Gata4 knockdown (green) or both (orange).

shown in Figure 59. To demonstrate additional features of the CARIN user interface, Figure 60 A-D shows a selection of further example networks.

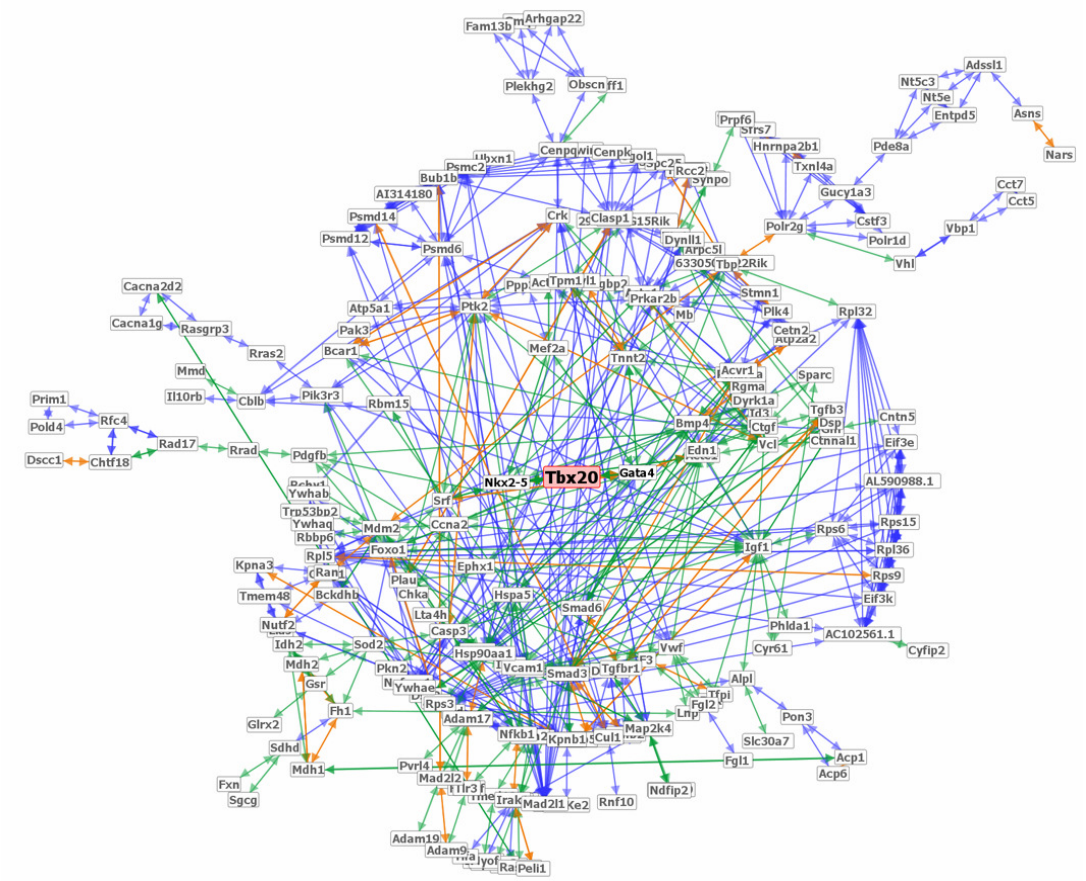

Figure 59: CARIN Example Session (4)
Network of genes that are connected to Tbx20 via STRING experiments and database or textmining associations and that are differentially expressed in the knockdown of at least one of Gata4, Mef2a, Nkx2.5 and Srf.

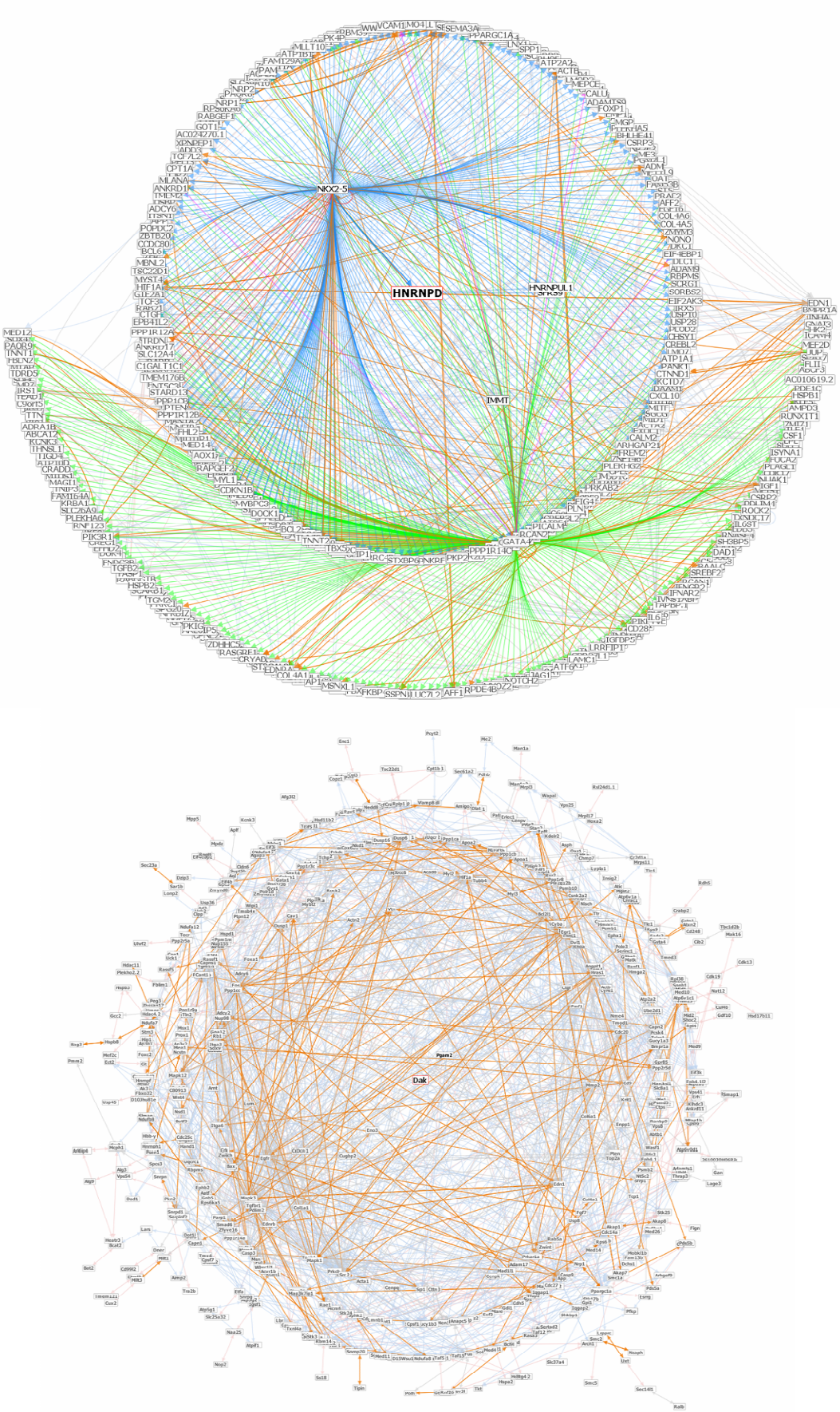

Figure 60: Further CARIN example networks (1)
(top) Gata4 and Nkx2.5 direct target genes measured using ChIP-on-chip with curved edges. **(bottom)** STRING and Anni associations for genes differentially expressed in Tbx18 heterozygous knockout.

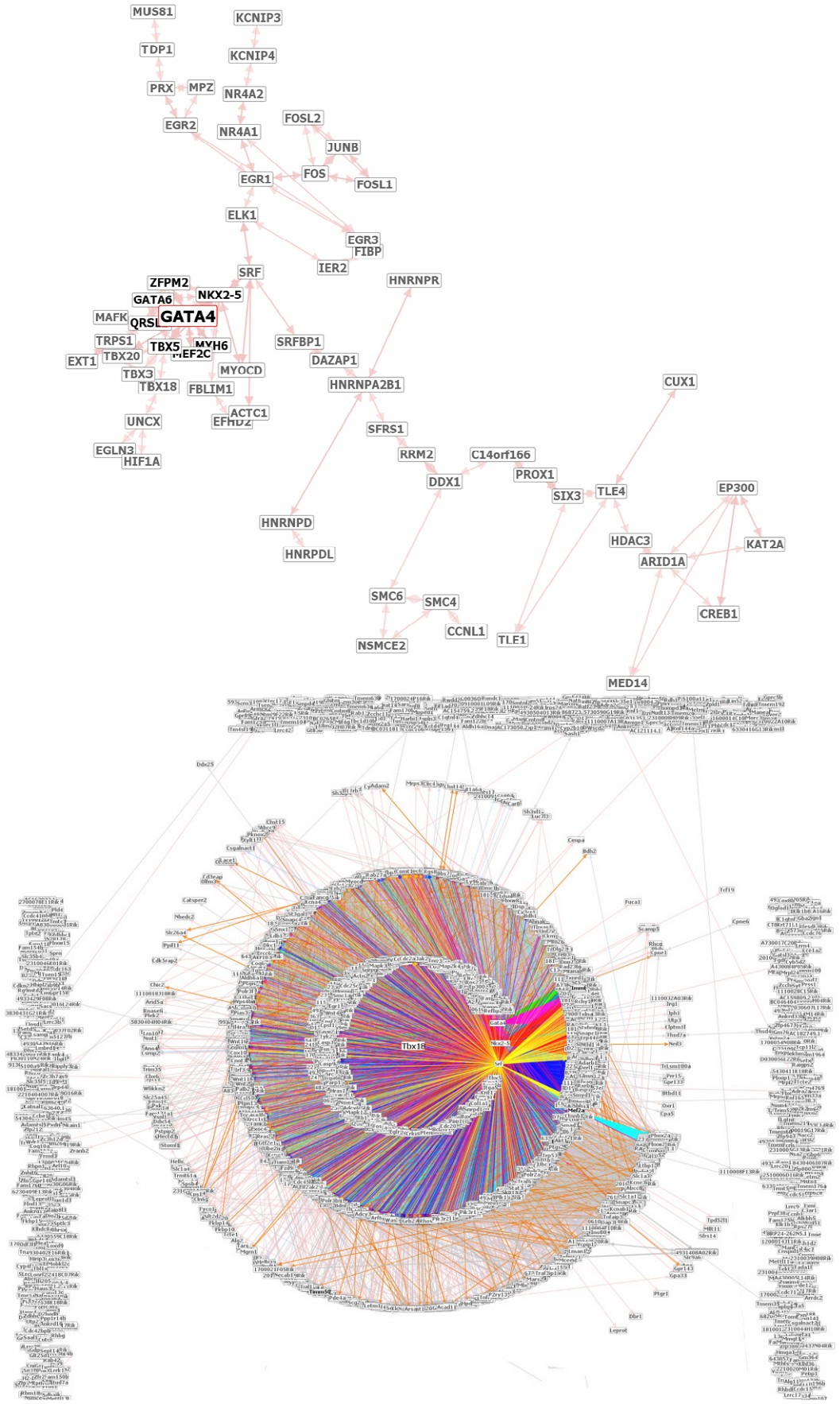

Figure 61: Further CARIN example networks (2)
(top) Subset of direct Srf targets connected to Gata4 through Anni textmining associations. **(bottom)** All annotated genes and interactions currently available in CARIN.

4. Discussion

This works represents a systematic *in vivo* analysis of important regulatory factors of the cardiac transcription network comprising DNA-binding transcription factors as well as epigenetic regulation by histone modifications. Genome-wide binding of the key cardiac transcription factors Gata4, Mef2a, Nkx2.5 and Srf was analyzed in conjunction with functional consequences of RNAi mediated knockdown of the individual factors in cell culture, leading to new insights into individual binding behaviour and function of the analyzed factors. These data were combined with DNA occupancy of activating histone modifications, revealing important regulatory dependencies that were confirmed and further analysed in a time-series of cardiac maturation in mouse around birth. In a final step, resulting transcription networks were studied and extended based on expression profile disturbances in patient with congenital heart disease. The study indicates a high complexity of the cardiac regulatory network, which is regulated on many levels comprising co-binding of transcription factors as well as the importance of accompanying histone modifications for TF target gene activation. This was elucidated for Srf and H3ac, which were studied in detail, indicating repeated buffering of the regulatory circuits but also high variability on the single gene level. Further, genes previously not shown to be linked to congenital heart disease were discovered to have an altered expression profile in a patient sharing similar disease phenotypes and functional upstream regulators were predicted.

In human and mouse ~2,000 transcription factors, more than 100 different modifications of histone residues and a large number of post-transcriptional regulators modulate the mRNA profile corresponding to 20,000-25,000 genes. A so far unknown fraction of these is important for correct heart development and function. Due to a large number of successful studies, major insights have been gained into the regulation of the transcriptional process by DNA-binding transcription factors and their modulators.[264-266] More recently, the roles of histone modifications in establishing and maintaining the chromatin status and their function as protein interaction partners have been discovered.[107,223,267] However, we lack data showing the interaction between these levels of regulation. Initial insights were obtained by focusing on each level and factor independently. Though, the goal of this thesis was a combinatorial analysis of the different regulatory mechanisms that drive correct cardiac gene expression. While it was long thought that transcription factors are the main driving force results of this study favor a comparable impact with a high degree of interdependency leading to a fine-tuned balance.

Using ChIP-chip, the transcription network driven by Gata4, Mef2a, Nkx2.5 and Srf was investigated. While these factors have already been analyzed in other studies, the binding of transcription factors is known to be cell type specific and this is the first study that analyzed all these TFs in the same cell type. Careful normalizing of the experimental data was performed to remove systematic experimental biases using a variance stabilizing normalization method. A peak calling algorithm based on a sliding window approach coupled to an empirically derived null distribution was applied to determine enriched binding sites for the individual factors in ChIP, revealing a large number of known as well as new targets for each factor, yet with large differences in their actual number. These results are in line with the current knowledge about the individual factors, *e.g.* Srf is known to have many targets as it is important not only in heart development but also in the general regulation of muscle cells.[268] GO term enrichment analysis[175] incorporating the GO topology revealed good agreement with previous knowledge about the individual factors substantiating the implemented approach. For example, targets as well as differentially expressed genes of all analyzed factors had a significant overrepresentation for

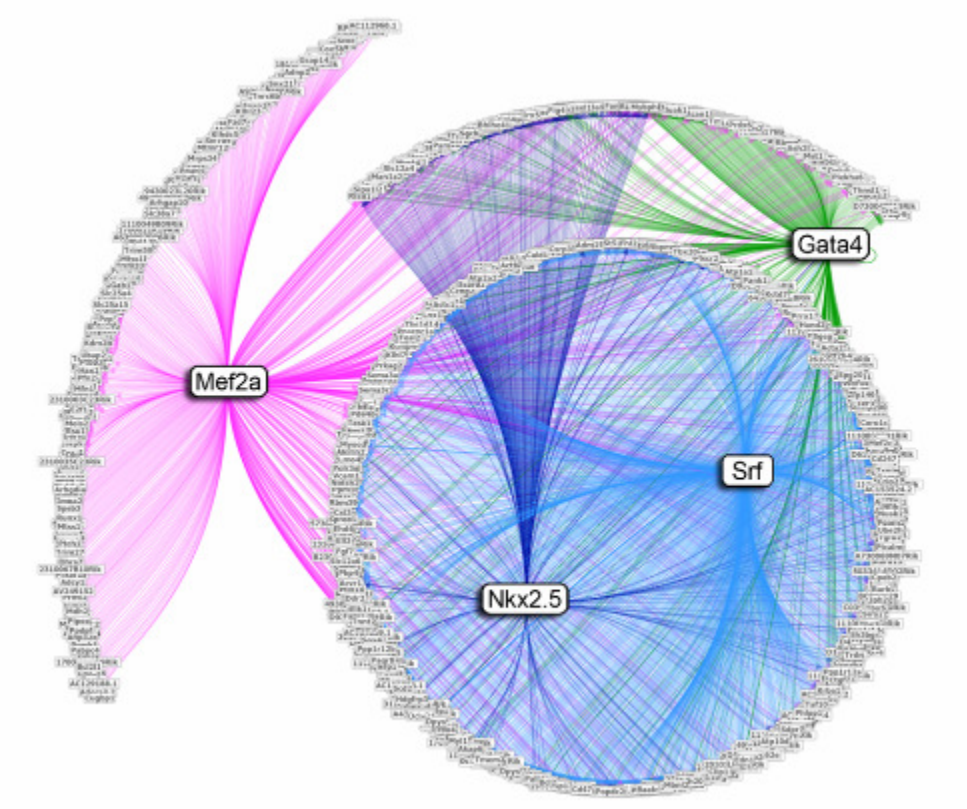

Figure 62: Network of direct targets gathered in ChIP-chip
The network shows direct targets of Gata4, Mef2a, Nkx2.5 and Srf gathered using ChIP-chip experiments in HL-1 cells. The high degree of overlapping targets is clearly revealed. This picture was produced using CARIN and later modified to enlarge the analyzed TFs.

heart and muscle related GO terms, like 'heart looping' and 'cell motility' for Nkx2.5 targets. Strikingly, mice with a partial knockout of Nkx2.5 show missing heart looping initiation[52] and reduced cell proliferation[269]. Further, genes annotated to 'bone mineralization' were also found significantly overrepresented in Nkx2.5 targets, eventually pointing to novel functions of this TF. To study the distribution of respective motifs in the found binding sites, ChIP-chip peaks were subsequently analyzed. Using both a *de novo* predictive approach as well as TFBS prediction with known PWMs resulted in the successful identification of binding motifs for Gata4, Mef2a and Nkx2.5 but failed to identify the TRANSFAC motif of Srf. Conversely, searching ChIP-chip peaks with the known CArG-box motif $CC(A/T)_6GG$.[53] resulted in a large number of positive peaks. It is suggestive that the low agreement between the motif search using TRANSFAC PWMs and the CArG-box motif might likely be due to an insufficient representation of the real Srf binding pattern by the existing PWMs. Many PWMs submitted to TRANSFAC rely on the '*Systematic Evolution of Ligands by Exponential Enrichment*' (SELEX) method,[270] which purifies and sequences binding sites of TFs. As the SELEX protocol contains very restrictive purification steps, many weaker binding sites will get lost, which was already suggested to lead to an insufficient representation of binding sites with intermediate affinity.[271]

Although physical interaction was shown for several of the analysed factors[272,273] only little was known about their extent of co-regulative binding *in vivo*. Therefore, combinatorial binding to the same promoter or in a close range was investigated, revealing a large number of common targets and binding sites. Figure 62 shows the full network of all bound targets indicating the high degree of

overlap. Comparing observed to expected number of pairwise targets, a positive correlation of binding sites was found for all factors except Srf, which also shared a large absolute number of targets with the other factors but was shown to have a more individual binding pattern. Interestingly, the lowest correlation (odds ratio ≈ 1/6) was found between Srf and Mef2a and these two factors were shown to bind to the same sites in a competitive manner to control gene expression.[274] Further, integrating binding of all factors simultaneously, a large number of almost 500 genes were shown to be bound by two or more of the analyzed factors, whereof 85 genes showed binding sites for all four in a close range of less than 500 bp. These results indicated a high degree of complexity and points to a cooperative regulation of gene expression.

To determine the functional consequences of TF binding, siRNA knockdown experiments were carried out and compared to wildtype expression using array detection. To remove systematic biases the datasets were normalized using a combination of a spline function based on sample quantiles[114] and the median polish algorithm to reduce the sequence bias.[208] Differentially expressed genes were determined using the limma[275] method revealing a mainly but not exclusively activating function of all analyzed factors. In line with other studies[276-278] it was found that most of the differentially expressed genes in the siRNA experiments were indirect targets of the respective transcription factor. *Vice versa*, many direct targets gathered in ChIP-chip did not show significantly altered expression in the respective siRNA knockdown. Potential reasons have been widely discussed in the literature.[265] The most prominent explanations given are TF binding in a poised state,[279] insufficient knockdowns[280] or buffering by redundant paralogs.[281] For example, members of the Mef transcription factor family were already postulated to buffer each others dysfunction.[205,282,283] In line with this, Mef2a was found to have the lowest number of differentially expressed genes in its knockdown. As an additional explanation, this study suggests a buffering effect by co-regulative binding also of non-paralogs, like the four investigates transcription factors. In accordance, those genes that were bound by multiple transcription factors in ChIP-chip were found to be significantly less likely differentially expressed in their respective siRNA knockdown. Likewise, factors with a high number of common targets showed only a low number of commonly differentially expressed genes in their knockdown.

To explore the influence of histone modifications as an epigenetic mechanism to modulate gene expression, the transcription factor binding data was analyzed in the context of co-occurring activating histone modification marks of H3ac, H3K4me2, H3K4me3 and H4ac.[16] It was found that ~80% of the observed transcription factor binding events were additionally marked by one or more of these histone modifications, whereas in a randomized situation only 23% are expected to co-occur. These observations are consistent with others studies which observed that regulatory binding sites are frequently marked by histone modifications.[225,284] Consequently it was investigated whether the presence of any of these histone modifications had an influence on the expression levels of direct target genes and a significant dependency was found for Gata4, Nkx2.5 as well as Srf but not for Mef2a. The fact that no positive effect of accompanying histone modifications on gene expression was found for Mef2a might correlate with the observation that Mef2a interacts with both HATs as well as HDACs, thereby acting as a platform to respond to both positive and negative transcriptional signals.[285,286]

Using linear modeling techniques, H3ac was revealed as the only of the investigated histone modification that modulates Srf target gene activation. Further, using ChIP-seq experiments followed by read mapping and peak calling based on a negative binomial distribution, this interaction was confirmed in a genome-wide manner. Incorporating results of the siRNA knockdown of Srf, it was

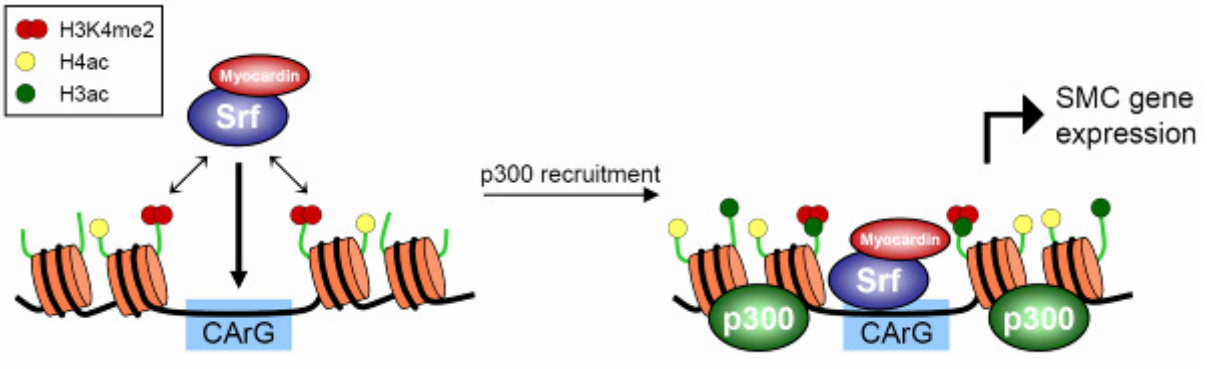

Figure 63: Gene activation by Srf, p300 and histone modifications
Model for smooth muscle cell (SMC) gene activation according to McDonald and Owens[230]. H4ac and H3K4me2 facilitate the binding of the Srf/Myocardin complex to regulatory CArG boxes. The complex then recruits p300, leading to further histone acetylation and active gene expression.

shown that Srf target genes additionally marked by H3ac preserved their high expression even after the cellular abundance of Srf was diminished, pointing to a buffering effect of the acetylation marks. Based on these findings, time-series experiments were designed to analyze the dependency between Srf and H3ac. Therefore ChIP-qPCR was performed for preselected regulatory sites gathered from mouse heart in three developmental stages of cardiac maturation around birth. In addition to the two aforementioned factors, the histone modification H3K4me2 as well as the HAT p300 were screened, which both were associated to Srf dependent gene activation in other studies.[230,231] After data normalization, correlated changes between subsequent stages were investigated, indicating high to modest dependencies between the individual factors. While a high correlation between changes in the HAT p300 and acetylation level was expected, the link between the other factors is not straightforward. Interestingly, the highest correlation was found between changes in H3ac and H3K4me2 levels. This supports findings in human CD4+T cells which revealed that histones in promoters that showed initial H3K4 methylation were acetylated after treatment with Trichostatin A, an HDAC inhibitor, while those promoters that did not had initial H3K4 methylation remained largely unacetylated.[222] The coupling between the histone tail methylation and acetylation might either be achieved through the recruitment of ATP-dependent chromatin remodeling complexes which open the chromatin to allow other histone modification enzymes to bind[227,228] or by direct H3K4me2 recognition from histone acetylation complexes.[229] It was further suggested that H3K4 methylation of nucleosome particles can increase the efficiency of acetylation by p300.[287] In line with this, we found significant correlation between changes in H3K4me2 and p300 enrichment.

The coupling of p300 to H3K4me2 and finally to H3ac might likely be established through tissue specific transcription factors. Following a recently suggested model by McDonald and Owens,[230] gene activation by Srf is performed in a step-wise procedure (Figure 63). In order to bind its CArG-box motif, initial histone modifications like H3K4me2 and H4ac must be present. H4ac thereby leads to an open chromatin state facilitating genomic access while H3K4me2 provides a docking site to the Srf/Myocardin complex. Myocardin, one of the most potent co-activator of Srf, belongs to the SAP domain family of nuclear proteins and is exclusively expressed in smooth muscle cells and cardiomyocytes. The binding of the Srf/Myocardin complex recruits further transcription factors like p300 and other members of the basal transcriptional machinery[62,221] via the powerful C-term transcriptional activation domain. This fully activates the gene expression program. Consistent with this model, lower albeit significant correlations were found between changes in Srf enrichment and the

two histone modifications. However, the correlation between p300 and Srf, while still remaining significant, was the lowest found in this analysis and was even lower than the correlation between Srf and H3ac. In addition, linear models that were built to estimate the influence of both Srf as well as p300 either in a qualitative or a quantitative way revealed a significant influence for both factors. Besides, the correlation between Srf and p300 was completely lost if only the two postnatal stages were taken into account. Taken together, these findings contradict with the proposed model of p300 recruitment by Srf but rather point to the existence of different factors that link Srf-binding to histone 3 acetylation.

The observed impact of histone 3 acetylation on the activating potential of transcription factors as exemplarily shown for Srf underlines the beneficial effects seen for HDAC inhibitors for a variety of disease states.[288] Results from this study favor the view that buffering by co-binding transcription factors as well as modulation of the histone modification status might be a plausible explanation for incomplete penetrance or phenotypic diversity as frequently observed in mouse models with identical genetic background or in human disease such as congenital heart disease. Here, a distinct gene mutation can lead to a broad portfolio of phenotypes, such as mutations in Cited-2 which lead to various cardiac malformations including atrial and ventricular septal defects.[289,290] To increase the current knowledge about transcription networks and their breakdown in heart disease, gene expression data for a selected set of cardiac relevant genes was analyzed in a large number of patients with a broad range of congenital heart disease. Using hierarchical clustering to compress the partially overlapping phenotype characteristics meta-phenotypic subgroups of patients were defined and linear modeling was used to determine phenotype specific gene expression changes. Thereby, specific molecular portraits of each meta-phenotype could be revealed with some genes that were specifically deregulated in only a single group but also those that were deregulated in more than one meta-phenotype group. Two promising examples were TBX20 and DPF3, which were both found to be up regulated in patients with Tetralogy of Fallot, one of the most common congenital heart disease, which has been linked to a number of genetic, epigenetic and environmental disturbances.[291] Of these, TBX20 was found to be specifically associated to patients with TOF and main pulmonary artery abnormalities, while DPF3 was associated to the same cluster but also to patient with TOF and a left-to-right shunt across the interventricular septum. Based on these findings two follow-up studies were conducted by members of our own group investigating TBX20 and DPF3 further (Hammer *et al.*[292] and Lange *et al.*[107]). Interestingly, *DPF3* was found to be a 'reader' of histone modifications and might provide a tissue specific anchor of the overall histone remodeling machinery, once more indicating the high interdependency between the individual regulatory levels.

An integrative approach was used to construct human regulatory transcriptional networks based on correlated gene expression and optimized prediction of transcription factor binding sites. Groups of genes showing a highly correlated expression pattern over all analyzed patients were used to reveal likely co-regulated genes. To predict common regulators, cis-regulatory elements present in the promoters of all correlated genes were determined using TFBS prediction that was initially optimized based on the ChIP-chip dataset to ensure the best agreement with experimental results. Resulting transcription networks were confirmed using literature as well as an additional affinity based transcription factor prediction. The implemented network prediction approach was able to shed light on a regulatory subnetworks driving TBX20 expression, which revealed *NKX2.5* and *GATA4* as potential regulators. TBX20 was not only shown to be associated to TOF in this study but is of particular interest as mutations in the TBX20 gene or changes in *TBX20* levels are in general

associated with severe congenital malformations both in animal models[293] and in patients.[294] However, the regulation of *TBX20* was not well known so far and the only described signaling molecule upstream of *TBX20* is BMP2.[295] Yet recently, studies in our own lab could show that TFAP2C, which was not analyzed in this study, is a direct regulator of TBX20.[292] Identification of NKX2.5 and GATA4 as common regulators reveals them as interesting candidate genes to be responsible for the transcription pattern of this phenotype cluster. A causative connection is suggestive and mutations in both TFs have already been linked to TOF.[79,296] Integrating the results with the cell culture experiments, *Tbx20* was in addition found to be a direct target of Mef2a and Srf in ChIP-chip. Further, siRNA knockdown experiments in all four TFs showed significant reduction of *Tbx20* mRNA level. These results demonstrate that binding by the four factors as found in ChIP-chip and partially predicted by the network approach is indeed functional and activates *Tbx20* expression. At the same time, the integration of the bioinformatic prediction provides an advantage over the use of only the ChIP experiments, as further potential modulators of TBX20 expression, like *HAND2* and *SMAD4* were revealed, which can now be tested using additional experiments.

Results of this study highly support a combinatorial nature of transcriptional regulation with a high degree of interdependency that is carefully orchestrated to regulate correct temporal and spatial expression of every gene to establish cellular function. In addition, as suggested by the performed knockdown experiments, these mechanisms are repeatedly buffered to maintain gene expression even in the case of disturbed regulators. As an example, Srf target gene activation was shown to be highly dependent on histone modifications and strikingly, histone modifying enzymes represent an important group of direct downstream targets of Srf as found in ChIP-chip/seq. *E.g.* the histone demethylases containing a Jumonji domain such as Jmjd1c, Jmjd2b, Jmjd3, Jmjd4 and Jmjd5 were all found to be direct Srf targets. Finally, transcription factors like DPF3 have the power to 'read' single histone modifications and suggestively even a histone code, leading to a tissue specific anchoring of the epigenetic machinery. Thus, analyzing the influence of individual regulators like transcription factors or histone modifications on target gene expression levels in a genome-wide manner can reveal the overall influence of each of these factors in general. However, on the single-gene level, a high variability in the regulatory dependencies was found. While the four TFs analyzed in ChIP-chip were shown to frequently bind together, many target genes exist that are only regulated by a subset of these. Further, manual analysis of target genes assigned to region studied in the ChIP-qPCR time series revealed highly dynamic sets of regulators that co-operatively work to regulate correct temporal transcription. The analysis of these dynamics as well as buffering mechanisms pose a great challenge on future analyses of regulatory networks.

As this study integrated a number of complementary experiments, bioinformatic challenges were manifold. One of the most important tasks in the analysis of high-throughput datasets is a careful normalization of the data to remove systematic biases. Though, the performance of different normalization methods can highly depends on the analyzed dataset.[297,298] In this study, this was demonstrated for the variance stabilization normalization method, which performed very well for the ChIP-chip data but lead to less significant results in the analysis of the siRNA knockdown expression data. As this was supposed to be dependent on missed parametric assumptions, it was in that case replaced by the non-parametric qspline normalization method, which led to higher agreement between replicated experiments and therefore enabled the detection of differentially expressed genes. Further, in case of ChIP-chip data, it has been suggested that probe sequence-specific normalization should be performed as it is usually done for gene expression microarrays. Johnson *et al.*[299,300] introduced the

first normalization models for ChIP-chip data that incorporated probe sequence composition. More recently, Chung *et al.*[301] showed that the enhanced ability of these methods to detect enriched DNA fragments was not based on the sequence dependent normalization but only on their implemented peak calling procedure. In line with their finding, no need for an additional sequence-specific normalization could be revealed in this study, as no prominent sequence bias could be detected after vsn normalization.

Due to their complex nature, the interpretation of ChIP-chip experiments was one of the most demanding bioinformatic challenges. So far, no clearly accepted consensus has been established how the ChIP-chip signal of real binding sites has to be distinguished from background, leading to the implementation and utilization of many different ChIP-chip peak calling methods today.[154,158,299,302-304] This problem is aggravate by the fact that per se only little true binding sites are known for each transcription factor, making empirical benchmarking of different methods very challenging.[305] Taking advantage of the single-tailed nature of the data, which is caused by specific enrichment but not specific depletion of genomic fragments,[154] the technique implemented in this study to call significantly enriched binding sites was developed hand in hand with the existing experimental data to ensure high accuracy of found peaks. Furthermore, results were justified with existing knowledge about the individual transcription factors from heart and muscle cells and partially confirmed for known target genes using separate experiments. As an substantiating fact, GO term enrichment analysis revealed significantly enriched functional annotations of the predicted targets that were in line with existing knowledge about the individual factors. In addition to the detection by microarrays, this study also incorporated the more recently developed next-generation sequencing technique to detect ChIP enrichment. ChIP-seq poses different requirements on its analysis, which rendered a reuse of the implemented peak finding implemented for ChIP-chip impossible. Instead, a peak calling algorithms tailored to the needs of ChIP-seq data was selected that implements the negative binomial distribution as this was shown to be more accurate than earlier approaches.[94] Interestingly, while both methods aim to measure the same binding sites, large different results of peak positions were found in the comparison of Srf ChIP-chip against ChIP-seq data. The reason for this difference, which was not only found in this study but also in a number of others, is so far unknown. Yet a number of possible explanations have been given in the literature, comprising differences in specificity and sensitivity and resolution of techniques.[93,94,219,220] Further, it is possible that the differences in detected binding sites aren't based on the different experimental techniques but the algorithmic approaches used for peak calling. Though, what disfavours this explanation was the high overlap of 91% that was observed for H3ac. Nevertheless, analyzing the ChIP-seq data in the same way as the ChIP-chip data, the same overall results were gained.

To predict regulatory networks in human patient, gene expression analysis was coupled to the prediction of *cis*-regulatory elements. To increase the signal-to-noise ratio frequently found especially in the promoters of higher vertebrates,[306] the prediction was restrained to tightly correlated and therefore likely co-regulated genes, using the Pearson correlation coefficient as the measure of statistical dependency. While it has been used in a large number of studies,[121-123] Pearson correlation measures only the linear dependency between variables and was proposed to miss important non-linear dependencies present in gene expression data.[124] Therefore, mutual information based on density estimation[131] was implemented as a measure of linear and non-linear dependencies in this study. However, comparing mutual information to Pearson correlation revealed no significant fraction of non-linear dependencies in the gene expression data. The same result was found in a more general

study comparing these two measures by Steuer *et al.*[127] Further, the MATCH algorithm was used to predict TFBS in the promoter of correlated genes. To derive the best parameters for the prediction process, meaning the length of the used promoter as well as the level of conservation, parameter optimization was performed incorporating known binding sites from the analyzed ChIP-chip data of Gata4, Mef2a and Nkx2.5 and the resulting parameters were applied also for the other TFs. This parameter optimization process indicated a promoter length of 1250 bp upstream and 500 bp downstream to be a good balance between too long promoters, resulting in a low signal-to-noise level, and too short promoters, which would miss important binding sites found in ChIP. Interestingly, the length of the optimal downstream region was found to be quite long, indicating a large number of binding sites potentially in intronic or even exonic regions. This high number of downstream functional elements was also revealed in an analysis performed by the ENCODE consortium.[264] In addition, the comparison with ChIP-chip peaks revealed that while the use of conservation information will highly reduce the sequence's noise ratio, indicated by a higher fraction of true predictions, requiring a to high conservation level will result in the loss of many functional binding sites, as revealed by the lower fraction of predicted ChIP peaks. This will especially be true for those sites which represent species-specific regulatory elements, as these will not be conserved. The results of the parameter optimization process revealed a minimal nucleotide conservation level of 60% in a 100 bp window to be the optimal choice for the analyzed dataset. This high variability in TF binding events even between very closely related species has been demonstrated in a number of animals, including different yeast[307] and drosophila[308] strains as well as between mouse and human.[309] Comparing the results from the prediction of single TFBS using MATCH with the affinity-based approach implemented in TRAP, a large number of predicted regulatory dependencies were found using both tools. This is not surprising, as both methods rely on the same binding motifs and a DNA sites that closely resemble the TF binding motif will lead to a prediction in MATCH as well as a high affinity in TRAP. Yet, the affinity approach yields the possibility to integrate the contribution of also weak binding sites, thereby providing the chance of detecting regulatory implications that will not be visible to a threshold based prediction such as MATCH. However, as shown in the case of NKX2-5, which had predicted and ChIP-validated regulatory binding sites in the promoters of four highly correlated genes but failed to be assigned as a high affinity regulator using TRAP, the highest affinity prediction as well as TFBS prediction in general does not always reflect biological binding known from literature or identified in ChIP.

While the approach to build regulatory networks by combining correlated gene expression with TFBS prediction has been implemented in a number of studies it is only one of many possible. Another very popular method is Bayesian network learning[310] for single state analysis or Dynamic Bayesian network learning for time-series data.[311] They are highly suggestive for modeling of regulatory networks as they provide indirect causality between the nodes of the estimated network, which can be translated into regulatory dependencies. Though, to reliably estimate the underlying probability density distributions of the regulatory network these methods require a large number of data measurements, which is limited by the amount of available material especially when analyzing human samples and was therefore not suitable for the analysis. Another promising approach for a joined analysis of phenotype and genotype data would be biclustering.[312] While normal clustering depends on exclusively rows or columns, or patients or genes, a biclustering approach simultaneously clusters both dimensions. Especially in respect to the mainly overlapping phenotypes of the dataset, biclustering would allow a clustering of the underlying genes without a fixed patient phenotype

grouping and could lead to the uncovering of further gene regulations, which might have been missed due to the hierarchical clustering approach. Indeed, biclustering was used to analyze the dataset. However, it was revealed that the small number of 42 genes was insufficient to drive meaningful results and the analysis was therefore excluded from this thesis.

Different to bioinformatic analyses of data of only a single experimental type, *e.g.* gene expression measurements, this study presents an integrative approach incorporating annotations gathered using multiple experimental techniques including the screening of *in vivo* transcription factor binding, gene expression measurements in wildtype and knockdown and bioinformatic predictions of regulatory dependencies. A number of similar integrative approaches have been published comprising the analysis of the mitochondrial proteome,[313] the hepatitis C[314] and influenza[315] viruses, miRNA regulatory circuits,[316] genetic interactions in yeast[317] and E. coli,[318] or the prediction of drug targets in human[319,320] amongst many others. Bioinformatic analysis is most vital for all of these studies. The integration of many complementary techniques yields the power to derive more comprehensive insights than each experiment individually could do by assaying varying aspects of complex biological processes, whether the focus of the study is a single-cell organism like yeast or a highly complex organ like the human heart. Thus, a combination of different approaches as done in this study will lead to more significant results in the light of biological authenticity. Further, this study presents one of the first analyses that investigated combinatorial TF regulation in cardiac transcription networks on a genome-wide scale systematically incorporating the influence of histone modifications. However, such genome-wide cardiovascular datasets are rare and often hard to relate as they have been gathered using different species and different platforms. In an attempt to collect existing datasets, the CArdiovascular Regulatory INteraction database was implemented, which has become the dissemination database for the EU project 'HeartRepair'. Given its structure which integrates data gathered using different experimental techniques and from different organisms as well as its sophisticated visual user interface, which incorporates the possibility to filter and highlight subgroups of genes, it is meant as a communication platform for researchers of different fields to conjointly work on questions related to heart development, function and disease.

Yet, a number of questions raised in this study still remain unanswered: given the predicted co-regulation of the four transcription factors analyzed in the cell culture experiments, it is still unclear how they interact and in what quantities this interaction is required for correct gene regulation or if it is merely a way to dynamically buffer changes in the enrichment of individual factors. Further, it was revealed that most of the analyzed factors require the co-occurrence of additional histone modifications, yet only a small number of all known modifications was investigated. In addition, for most of the interaction between transcription factor and histone modifications it is unknown how their coupling is established and, with the exception of a small fraction of histone modifications, how their presence influences gene expression. While results from this study substantiated the supposed coupling of Srf and H3ac, the existing model that predicts a coupling through the histone acetyltransferase p300 could not be confirmed. Finally, the large number of children born with congenital heart disease still poses a severe challenge on today's research. So far, the correlated interaction of factors in transcriptional networks is largely unknown, especially in respect to their dynamic behavior given long temporal processes such as development. This lack of knowledge is exemplified by the low number of patients with congenital heart disease that carry mutations within the few identified causal human genes.[321] Given previous genetic studies, it is assumed that most congenital heart defects have a multi-genetic and multi-factorial basis with partially overlapping

genotypic and phenotypic manifestations. Results from this thesis and other recent studies demonstrate that histone modifications are likely contributors to heart disease and must therefore be integrated as additional layers of transcriptional regulation.[322-325]

To systematically investigate the genetics of congenital heart disease, next-generation sequencing of RNA in patients is likely the method of choice. In combination with family studies or population-wide sequence variation analysis it provides the basis for high-throughput analyses of expression level and sequence mutational causes for heart and other congenital disease. In addition, mRNA sequencing allows the identification and quantification of new splice variants. A number of genetic diseases have already been linked to incorrect mRNA-processing[326] and it has been suggested that at least 15% of all single base-pair mutations causing human genetic diseases result in pre-mRNA splicing defects.[327] Some of these mutations can create new splice sites, while others weaken regular splice sites, thereby leading to the recognition of nearby pre-existing cryptic splice sites. Further, mutations in exon sequences were found that disturb splicing protein binding and lead to an exclusion of the appropriate exon from the final mRNA. And just recently, these processes have been further complicated by the finding that histone modification can directly regulate alternative splicing of pre-mRNAs.[328]

Two other important regulatory mechanisms which influence the abundance of transcripts, namely the presence of DNA-methylation sites in promoter regions and the post-transcriptional regulation by miRNAs, have not been considered in this analysis. Nevertheless, they were shown to have implications on heart development and disease. Especially miRNAs are the topic of many recent studies. It has been described that miRNA expression profiles change during cardiac development and just recently it has been shown that a considerable amount of proteins are deregulated during heart failure by microRNA activation and/or silencing.[329] Interestingly, an impressive similarity has been found between the miRNA expression patterns occurring in human failing heart and 14-weeks-old fetuses, like the up regulation of miR-21, miR-29b and miR-210 and the down regulation of miR-30, miR-182 and miR-526.[330] Gene expression analysis coupled to miRNA target prediction revealed that most up regulated genes were characterized by the presence of a significant number of predicted binding sites for down regulated miRNAs and *vice versa*.[329] The expression of miR-1 was shown to have an impact on embryo lethality in mice and thus its dysregulation might very likely result in congenital heart disease in humans.[39] Again, genome-wide miRNA sequencing approaches preferably in conjunction with mRNA sequencing might very likely underline this proposed implication. Studies correlating these two approaches are recently ongoing in our group. As an interesting fact, new chemically engineered oligonucleotides, so-called ‘antagomirs’ have been proven to specifically and efficiently silence miRNA in mice, implicating their use as potential therapeutic substances.[331,332] In addition, histone-modifying enzymes are already utilized as therapeutic substances especially in the field of cancer treatment.

Combining the findings from genome-wide studies, time-series analysis of correlated binding changes and the prediction of cardiac regulatory network and disease-associated molecular portraits as done in this thesis can therefore suggest novel cardiac regulatory cascades and might even point out strategies for therapeutic treatment of pathological cardiac growth, remodeling and heart failure.

5. Bibliography

1. R. D. Kornberg, Chromatin structure: a repeating unit of histones and DNA, *Science (New York, N.Y*, 1974, **184**, 868-871.
2. R. D. Hawkins and B. Ren, Genome-wide location analysis: insights on transcriptional regulation, *Human molecular genetics*, 2006, **15 Spec No 1**, R1-7.
3. L. O. Barrera and B. Ren, The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding, *Current opinion in cell biology*, 2006, **18**, 291-298.
4. M. Kishimoto, R. Fujiki, S. Takezawa, Y. Sasaki, T. Nakamura, et al., Nuclear receptor mediated gene regulation through chromatin remodeling and histone modifications, *Endocrine journal*, 2006, **53**, 157-172.
5. B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, et al., Genomic maps and comparative analysis of histone modifications in human and mouse, *Cell*, 2005, **120**, 169-181.
6. H. R. Chung and M. Vingron, Sequence-dependent nucleosome positioning, *Journal of molecular biology*, 2009, **386**, 1411-1422.
7. I. P. Ioshikhes, I. Albert, S. J. Zanton and B. F. Pugh, Nucleosome positions predicted through comparative genomics, *Nature genetics*, 2006, **38**, 1210-1215.
8. G. C. Yuan and J. S. Liu, Genomic sequence is highly predictive of local nucleosome depletion, *PLoS computational biology*, 2008, **4**, e13.
9. E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, et al., A genomic code for nucleosome positioning, *Nature*, 2006, **442**, 772-778.
10. H. Zhang, D. N. Roberts and B. R. Cairns, Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss, *Cell*, 2005, **123**, 219-231.
11. S. L. Berger, The complex language of chromatin regulation during transcription, *Nature*, 2007, **447**, 407-412.
12. B. D. Strahl and C. D. Allis, The language of covalent histone modifications, *Nature*, 2000, **403**, 41-45.
13. V. G. Allfrey, R. Faulkner and A. E. Mirsky, Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis, *Proceedings of the National Academy of Sciences of the United States of America*, 1964, **51**, 786-794.
14. Y. Bao and X. Shen, SnapShot: chromatin remodeling complexes, *Cell*, 2007, **129**, 632.
15. N. Dillon and R. Festenstein, Unravelling heterochromatin: competition between positive and negative factors regulates accessibility, *Trends Genet*, 2002, **18**, 252-258.
16. J. J. Fischer, J. Toedling, T. Krueger, M. Schueler, W. Huber, et al., Combinatorial effects of four histone modifications in transcription and differentiation, *Genomics*, 2008, **91**, 41-51.
17. R. Karlic, H. R. Chung, J. Lasserre, K. Vlahovicek and M. Vingron, Histone modification levels are predictive for gene expression, *Proceedings of the National Academy of Sciences of the United States of America*, 2010, **107**, 2926-2931.
18. S. R. Bhaumik, E. Smith and A. Shilatifard, Covalent modifications of histones during development and disease pathogenesis, *Nature structural & molecular biology*, 2007, **14**, 1008-1016.
19. K. L. Tucker, Methylated cytosine and the brain: a new base for neuroscience, *Neuron*, 2001, **30**, 649-652.
20. B. H. Ramsahoye, D. Biniszkiewicz, F. Lyko, V. Clark, A. P. Bird, et al., Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA

methyltransferase 3a, *Proceedings of the National Academy of Sciences of the United States of America*, 2000, **97**, 5237-5242.
21. R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, et al., Human DNA methylomes at base resolution show widespread epigenomic differences, *Nature*, 2009, **462**, 315-322.
22. A. C. D'Alessio and M. Szyf, Epigenetic tete-a-tete: the bilateral relationship between chromatin modifications and DNA methylation, *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 2006, **84**, 463-476.
23. J. Winter, S. Jung, S. Keller, R. I. Gregory and S. Diederichs, Many roads to maturity: microRNA biogenesis pathways and their regulation, *Nature cell biology*, 2009, **11**, 228-234.
24. C. Cheadle, J. Fan, Y. S. Cho-Chung, T. Werner, J. Ray, et al., Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability, *BMC genomics*, 2005, **6**, 75.
25. D. P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell*, 2004, **116**, 281-297.
26. D. P. Bartel, MicroRNAs: target recognition and regulatory functions, *Cell*, 2009, **136**, 215-233.
27. M. Inui, G. Martello and S. Piccolo, MicroRNA control of signal transduction, *Nature reviews*, **11**, 252-263.
28. A. J. Pratt and I. J. MacRae, The RNA-induced silencing complex: a versatile gene-silencing machine, *The Journal of biological chemistry*, 2009, **284**, 17897-17901.
29. I. Bentwich, A. Avniel, Y. Karov, R. Aharonov, S. Gilad, et al., Identification of hundreds of conserved and nonconserved human microRNAs, *Nature genetics*, 2005, **37**, 766-770.
30. R. C. Friedman, K. K. Farh, C. B. Burge and D. P. Bartel, Most mammalian mRNAs are conserved targets of microRNAs, *Genome research*, 2009, **19**, 92-105.
31. J. Brennecke, A. Stark, R. B. Russell and S. M. Cohen, Principles of microRNA-target recognition, *PLoS biology*, 2005, **3**, e85.
32. L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, et al., Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs, *Nature*, 2005, **433**, 769-773.
33. Y. Tan, B. Zhang, T. Wu, G. Skogerbo, X. Zhu, et al., Transcriptional inhibiton of Hoxd4 expression by miRNA-10a in human breast cancer cells, *BMC molecular biology*, 2009, **10**, 12.
34. S. Gonzalez, D. G. Pisano and M. Serrano, Mechanistic principles of chromatin remodeling guided by siRNAs and miRNAs, *Cell cycle (Georgetown, Tex*, 2008, **7**, 2601-2608.
35. A. S. Yoo, B. T. Staahl, L. Chen and G. R. Crabtree, MicroRNA-mediated switching of chromatin-remodelling complexes in neural development, *Nature*, 2009, **460**, 642-646.
36. G. Chua, M. D. Robinson, Q. Morris and T. R. Hughes, Transcriptional networks: reverse-engineering gene regulation on a global scale, *Current opinion in microbiology*, 2004, **7**, 638-646.
37. S. Martin-Puig, Z. Wang and K. R. Chien, Lives of a heart cell: tracing the origins of cardiac progenitors, *Cell stem cell*, 2008, **2**, 320-331.
38. S. M. Wu, K. R. Chien and C. Mummery, Origins and fates of cardiovascular progenitor cells, *Cell*, 2008, **132**, 537-543.
39. B. G. Bruneau, The developmental genetics of congenital heart disease, *Nature*, 2008, **451**, 943-948.
40. Zoofari, *http://en.wikipedia.org/wiki/File:Heart_diagram_blood_flow_en.svg*, 2010.

41. K. K. Linask and J. W. Lash, Early heart development: dynamics of endocardial cell sorting suggests a common origin with cardiomyocytes, *Dev Dyn*, 1993, **196**, 62-69.
42. R. G. Kelly, Building the right ventricle, *Circulation research*, 2007, **100**, 943-945.
43. V. M. Christoffels, J. B. Burch and A. F. Moorman, Architectural plan for the heart: early patterning and delineation of the chambers and the nodes, *Trends in cardiovascular medicine*, 2004, **14**, 301-307.
44. M. H. Soonpaa, K. K. Kim, L. Pajak, M. Franklin and L. J. Field, Cardiomyocyte DNA synthesis and binucleation during murine development, *The American journal of physiology*, 1996, **271**, H2183-2189.
45. E. N. Olson, Gene regulatory networks in the evolution and development of the heart, *Science (New York, N.Y*, 2006, **313**, 1922-1927.
46. M. Heikinheimo, J. M. Scandrett and D. B. Wilson, Localization of transcription factor GATA-4 to regions of the mouse embryo involved in cardiac development, *Developmental biology*, 1994, **164**, 361-373.
47. J. D. Molkentin, Q. Lin, S. A. Duncan and E. N. Olson, Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis, *Genes & development*, 1997, **11**, 1061-1072.
48. K. Iida, K. Hidaka, M. Takeuchi, M. Nakayama, C. Yutani, et al., Expression of MEF2 genes during human cardiac development, *The Tohoku journal of experimental medicine*, 1999, **187**, 15-23.
49. S. L. Amacher, J. N. Buskin and S. D. Hauschka, Multiple regulatory elements contribute differentially to muscle creatine kinase enhancer activity in skeletal and cardiac muscle, *Molecular and cellular biology*, 1993, **13**, 2753-2764.
50. T. J. Lints, L. M. Parsons, L. Hartley, I. Lyons and R. P. Harvey, Nkx-2.5: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants, *Development (Cambridge, England)*, 1993, **119**, 419-431.
51. R. Bodmer, The gene tinman is required for specification of the heart and visceral muscles in Drosophila, *Development (Cambridge, England)*, 1993, **118**, 719-729.
52. I. Lyons, L. M. Parsons, L. Hartley, R. Li, J. E. Andrews, et al., Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5, *Genes & development*, 1995, **9**, 1654-1666.
53. J. M. Miano, X. Long and K. Fujiwara, Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus, *American journal of physiology*, 2007, **292**, C70-81.
54. N. S. Belaguli, L. A. Schildmeyer and R. J. Schwartz, Organization and myogenic restricted expression of the murine serum response factor gene. A role for autoregulation, *The Journal of biological chemistry*, 1997, **272**, 18222-18231.
55. D. Z. Wang and E. N. Olson, Control of smooth muscle development by the myocardin family of transcriptional coactivators, *Current opinion in genetics & development*, 2004, **14**, 558-566.
56. S. Arsenian, B. Weinhold, M. Oelgeschlager, U. Ruther and A. Nordheim, Serum response factor is essential for mesoderm formation during mouse embryogenesis, *The EMBO journal*, 1998, **17**, 6289-6299.
57. E. M. Small and P. A. Krieg, Transgenic analysis of the atrialnatriuretic factor (ANF) promoter: Nkx2-5 and GATA-4 binding sites are required for atrial specific expression of ANF, *Developmental biology*, 2003, **261**, 116-131.
58. S. Morin, F. Charron, L. Robitaille and M. Nemer, GATA-dependent recruitment of MEF2 proteins to target promoters, *The EMBO journal*, 2000, **19**, 2046-2055.
59. Y. Lee, T. Shioi, H. Kasahara, S. M. Jobe, R. J. Wiese, et al., The cardiac tissue-restricted homeobox protein Csx/Nkx2.5 physically associates with the zinc finger

protein GATA4 and cooperatively activates atrial natriuretic factor gene expression, *Molecular and cellular biology*, 1998, **18**, 3120-3129.

60. N. S. Belaguli, J. L. Sepulveda, V. Nigam, F. Charron, M. Nemer, et al., Cardiac tissue enriched factors serum response factor and GATA-4 are mutual coregulators, *Molecular and cellular biology*, 2000, **20**, 7550-7558.
61. B. G. Bruneau, Transcriptional regulation of vertebrate cardiac morphogenesis, *Circulation research*, 2002, **90**, 509-519.
62. D. Cao, Z. Wang, C. L. Zhang, J. Oh, W. Xing, et al., Modulation of smooth muscle gene expression by association of histone acetyltransferases and deacetylases with myocardin, *Molecular and cellular biology*, 2005, **25**, 364-376.
63. T. P. Yao, S. P. Oh, M. Fuchs, N. D. Zhou, L. E. Ch'ng, et al., Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300, *Cell*, 1998, **93**, 361-372.
64. T. Kouzarides, Chromatin modifications and their function, *Cell*, 2007, **128**, 693-705.
65. F. J. Davis, M. Gupta, B. Camoretti-Mercado, R. J. Schwartz and M. P. Gupta, Calcium/calmodulin-dependent protein kinase activates serum response factor transcription activity by its dissociation from histone deacetylase, HDAC4. Implications in cardiac muscle gene regulation during hypertrophy, *The Journal of biological chemistry*, 2003, **278**, 20047-20058.
66. J. I. Hoffman and S. Kaplan, The incidence of congenital heart disease, *Journal of the American College of Cardiology*, 2002, **39**, 1890-1900.
67. J. I. Hoffman, Incidence of congenital heart disease: II. Prenatal incidence, *Pediatric cardiology*, 1995, **16**, 155-165.
68. M. Nemer, Genetic insights into normal and abnormal heart development, *Cardiovasc Pathol*, 2008, **17**, 48-54.
69. R. H. Anderson and M. Tynan, Tetralogy of Fallot--a centennial review, *International journal of cardiology*, 1988, **21**, 219-232.
70. M. Ruiz, *http://en.wikipedia.org/wiki/File:Tetralogy_of_Fallot.svg*, 2006.
71. H. Matsson, J. Eason, C. S. Bookwalter, J. Klar, P. Gustavsson, et al., Alpha-cardiac actin mutations produce atrial septal defects, *Human molecular genetics*, 2008, **17**, 256-265.
72. V. Garg, I. S. Kathiriya, R. Barnes, M. K. Schluterman, I. N. King, et al., GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5, *Nature*, 2003, **424**, 443-447.
73. Z. A. Eldadah, A. Hamosh, N. J. Biery, R. A. Montgomery, M. Duke, et al., Familial Tetralogy of Fallot caused by mutation in the jagged1 gene, *Human molecular genetics*, 2001, **10**, 163-169.
74. Y. H. Ching, T. K. Ghosh, S. J. Cross, E. A. Packham, L. Honeyman, et al., Mutation in myosin heavy chain 6 causes atrial septal defect, *Nature genetics*, 2005, **37**, 423-428.
75. L. Zhu, R. Vranckx, P. Khau Van Kien, A. Lalande, N. Boisset, et al., Mutations in myosin heavy chain 11 cause a syndrome associating thoracic aortic aneurysm/aortic dissection and patent ductus arteriosus, *Nature genetics*, 2006, **38**, 343-349.
76. A. Megarbane, N. Salem, E. Stephan, R. Ashoush, D. Lenoir, et al., X-linked transposition of the great arteries and incomplete penetrance among males with a nonsense mutation in ZIC3, *Eur J Hum Genet*, 2000, **8**, 704-708.
77. I. C. Joziasse, J. J. van de Smagt, K. Smith, J. Bakkers, G. J. Sieswerda, et al., Genes in congenital heart disease: atrioventricular valve formation, *Basic research in cardiology*, 2008, **103**, 216-227.

78. B. Kaynak, A. von Heydebreck, S. Mebus, D. Seelow, S. Hennig, et al., Genome-wide array analysis of normal and malformed human hearts, *Circulation*, 2003, **107**, 2467-2474.
79. D. W. Benson, G. M. Silberbach, A. Kavanaugh-McHugh, C. Cottrill, Y. Zhang, et al., Mutations in the cardiac transcription factor NKX2.5 affect diverse cardiac developmental pathways, *The Journal of clinical investigation*, 1999, **104**, 1567-1573.
80. C. T. Basson, D. R. Bachinsky, R. C. Lin, T. Levi, J. A. Elkins, et al., Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome, *Nature genetics*, 1997, **15**, 30-35.
81. J. L. t. Hartman, B. Garvik and L. Hartwell, Principles for the buffering of genetic variation, *Science (New York, N.Y*, 2001, **291**, 1001-1004.
82. S. L. Rutherford, Between genotype and phenotype: protein chaperones and evolvability, *Nature reviews*, 2003, **4**, 263-274.
83. E. F. Zimmerman, Substance abuse in pregnancy: teratogenesis, *Pediatric annals*, 1991, **20**, 541-544, 546-547.
84. A. A. Starreveld-Zimmerman, W. J. van der Kolk, J. Elshove and H. Meinardi, Teratogenicity of antiepileptic drugs, *Clinical neurology and neurosurgery*, 1975, **77**, 81-95.
85. C. A. Loffredo, E. K. Silbergeld, C. Ferencz and J. Zhang, Association of transposition of the great arteries in infants with maternal exposures to herbicides and rodenticides, *American journal of epidemiology*, 2001, **153**, 529-536.
86. J. Bentham and S. Bhattacharya, Genetic mechanisms controlling cardiovascular development, *Annals of the New York Academy of Sciences*, 2008, **1123**, 10-19.
87. S. J. Cross, Y. H. Ching, Q. Y. Li, L. Armstrong-Buisseret, S. Spranger, et al., The mutation spectrum in Holt-Oram syndrome, *Journal of medical genetics*, 2000, **37**, 785-787.
88. T. K. Ghosh, E. A. Packham, A. J. Bonser, T. E. Robinson, S. J. Cross, et al., Characterization of the TBX5 binding site and analysis of mutations that cause Holt-Oram syndrome, *Human molecular genetics*, 2001, **10**, 1983-1994.
89. Y. Hiroi, S. Kudoh, K. Monzen, Y. Ikeda, Y. Yazaki, et al., Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation, *Nature genetics*, 2001, **28**, 276-280.
90. J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, et al., Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, *Genome biology*, 2002, **3**, RESEARCH0034.
91. P. Polak and E. Domany, Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes, *BMC genomics*, 2006, **7**, 133.
92. C. Schonbach, From masking repeats to identifying functional repeats in the mouse transcriptome, *Briefings in bioinformatics*, 2004, **5**, 107-117.
93. D. E. Schones and K. Zhao, Genome-wide approaches to studying chromatin modifications, *Nature reviews*, 2008, **9**, 179-191.
94. H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, et al., An integrated software system for analyzing ChIP-chip and ChIP-seq data, *Nature biotechnology*, 2008, **26**, 1293-1300.
95. A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, et al., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, *Genome research*, 2005, **15**, 1034-1050.
96. C. L. Liu, T. Kaplan, M. Kim, S. Buratowski, S. L. Schreiber, et al., Single-nucleosome mapping of histone modifications in S. cerevisiae, *PLoS biology*, 2005, **3**, e328.

97. D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, et al., Genome-wide map of nucleosome acetylation and methylation in yeast, *Cell*, 2005, **122**, 517-527.
98. B. E. Bernstein, E. L. Humphrey, R. L. Erlich, R. Schneider, P. Bouman, et al., Methylation of histone H3 Lys 4 in coding regions of active genes, *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**, 8695-8700.
99. D. Schubeler, D. M. MacAlpine, D. Scalzo, C. Wirbelauer, C. Kooperberg, et al., The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote, *Genes & development*, 2004, **18**, 1263-1271.
100. B. Kaynak, A. von Heydebreck, S. Mebus, D. Seelow, S. Hennig, et al., *Genome-wide array analysis of normal and malformed human hearts*, 2003.
101. A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, et al., A gene atlas of the mouse and human protein-encoding transcriptomes, *Proc. Natl. Acad. Sci. U S A*, 2004, **101**, 6062-6067.
102. R. Tabibiazar, R. A. Wagner, A. Liao and T. Quertermous, Transcriptional profiling of the heart reveals chamber-specific gene expression patterns, *Circ. Res.*, 2003, **93**, 1193-1201.
103. V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, et al., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic acids research*, 2006, **34**, D108-110.
104. I. Illumina, Sequencing Analysis Software - User Guide, 2008.
105. A. Hirschy, F. Schatzmann, E. Ehler and J. C. Perriard, Establishment of cardiac cytoarchitecture in the developing mouse heart, *Developmental biology*, 2006, **289**, 430-441.
106. T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, et al., Ensembl 2009, *Nucleic acids research*, 2009, **37**, D690-697.
107. M. Lange, B. Kaynak, U. B. Forster, M. Tonjes, J. J. Fischer, et al., Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex, *Genes & development*, 2008, **22**, 2370-2384.
108. D. Seelow, R. Galli, S. Mebus, H. P. Sperling, H. Lehrach, et al., d-matrix - database exploration, visualization and analysis, *BMC bioinformatics*, 2004, **5**, 168.
109. *R: A language and environment for statistical computing*, 2005.
110. R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, et al., Bioconductor: open software development for computational biology and bioinformatics, *Genome biology*, 2004, **5**, R80.
111. Å. Björck, *Numerical methods for least squares problems*, 1996.
112. J. J. Faraway, *Practical Regression and Anova using R*, 2002.
113. W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka and M. Vingron, Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics (Oxford, England)*, 2002, **18 Suppl 1**, S96-104.
114. C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, et al., A new non-linear normalization method for reducing variability in DNA microarray experiments, *Genome biology*, 2002, **3**, research0048.
115. W. Wu, N. Dave, G. C. Tseng, T. Richards, E. P. Xing, et al., Comparison of normalization methods for CodeLink Bioarray data, *BMC bioinformatics*, 2005, **6**, 309.
116. C. C. Barbacioru, Y. Wang, R. D. Canales, Y. A. Sun, D. N. Keys, et al., Effect of various normalization methods on Applied Biosystems expression array system data, *BMC bioinformatics*, 2006, **7**, 533.
117. T. S. Davison, C. D. Johnson and B. F. Andruss, Analyzing micro-RNA expression using microarrays, *Methods in enzymology*, 2006, **411**, 14-34.

118. D. M. Rocke and B. Durbin, A model for measurement error for gene expression arrays, *J Comput Biol*, 2001, **8**, 557-569.
119. R. Tibshirani, Variance stabilization and the bootstrap, *Biometrika*, 1988, 433--444.
120. W. Huber, A. von Heydebreck, H. Sueltmann, A. Poustka and M. Vingron, Parameter estimation for the calibration and variance stabilization of microarray data, *Statistical applications in genetics and molecular biology*, 2003, **2**, Article3.
121. M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences of the United States of America*, 1998, **95**, 14863-14868.
122. I. Lee, S. V. Date, A. T. Adai and E. M. Marcotte, A probabilistic functional network of yeast genes, *Science (New York, N.Y*, 2004, **306**, 1555-1558.
123. J. M. Stuart, E. Segal, D. Koller and S. K. Kim, A gene-coexpression network for global discovery of conserved genetic modules, *Science (New York, N.Y*, 2003, **302**, 249-255.
124. A. J. Butte and I. S. Kohane, Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, *Pacific Symposium on Biocomputing*, 2000, 418-429.
125. I. Priness, O. Maimon and I. Ben-Gal, Evaluation of gene-expression clustering via mutual information distance measure, *BMC bioinformatics*, 2007, **8**, 111.
126. C. E. Shannon and W. Weaver, *The mathematical theory of communication*, University of Illionis, Urbana, 1949.
127. R. Steuer, J. Kurths, C. O. Daub, J. Weise and J. Selbig, The mutual information: detecting and evaluating dependencies between variables, *Bioinformatics (Oxford, England)*, 2002, **18 Suppl 2**, S231-240.
128. G. S. Michaels, D. B. Carr, M. Askenazi, S. Fuhrman, X. Wen, et al., Cluster analysis and data visualization of large-scale gene expression data, *Pacific Symposium on Biocomputing*, 1998, 42-53.
129. R. Herwig, A. J. Poustka, C. Muller, C. Bull, H. Lehrach, et al., Large-scale clustering of cDNA-fingerprinting data, *Genome research*, 1999, **9**, 1093-1105.
130. S. Kullback, *Information theory and statistics*, Dover publ., Mineola (N.Y.), 1997.
131. Y. I. Moon, B. Rajagopalan and U. Lall, Estimation of mutual information using kernel density estimators, *Physical review*, 1995, **52**, 2318-2321.
132. T. Palomero, W. K. Lim, D. T. Odom, M. L. Sulis, P. J. Real, et al., NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth, *Proceedings of the National Academy of Sciences of the United States of America*, 2006, **103**, 18261-18266.
133. A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, et al., Reverse engineering cellular networks, *Nature protocols*, 2006, **1**, 662-671.
134. A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, et al., ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC bioinformatics*, 2006, **7 Suppl 1**, S7.
135. K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, et al., Reverse engineering of regulatory networks in human B cells, *Nature genetics*, 2005, **37**, 382-390.
136. D. C. Kim, X. Wang, C. R. Yang and J. Gao, Learning biological network using mutual information and conditional independence, *BMC bioinformatics*, **11 Suppl 3**, S9.
137. L. Wang, M. Montano, M. Rarick and P. Sebastiani, Conditional clustering of temporal expression profiles, *BMC bioinformatics*, 2008, **9**, 147.
138. S. Lloyd, Least squares quantization in PCM, *Information Theory, IEEE Transactions on*, 1982, **28**, 129-137.

139. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, et al., Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proceedings of the National Academy of Sciences of the United States of America*, 1999, **96**, 2907-2912.
140. J. Herrero, A. Valencia and J. Dopazo, A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics (Oxford, England)*, 2001, **17**, 126-136.
141. Y. Cheng and G. M. Church, Biclustering of expression data, *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB*, 2000, **8**, 93-103.
142. G. Getz, E. Levine and E. Domany, Coupled two-way clustering analysis of gene microarray data, *Proceedings of the National Academy of Sciences of the United States of America*, 2000, **97**, 12079-12084.
143. A. Ben-Dor, B. Chor, R. Karp and Z. Yakhini, Discovering local structure in gene expression data: the order-preserving submatrix problem, *J Comput Biol*, 2003, **10**, 373-384.
144. S. Bergmann, J. Ihmels and N. Barkai, Iterative signature algorithm for the analysis of large-scale gene expression data, *Phys Rev E Stat Nonlin Soft Matter Phys*, 2003, **67**, 031902.
145. J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, et al., Revealing modular organization in the yeast transcriptional network, *Nature genetics*, 2002, **31**, 370-377.
146. R. R. Delongchamp, J. F. Bowyer, J. J. Chen and R. L. Kodell, Multiple-testing strategy for analyzing cDNA array data on gene expression, *Biometrics*, 2004, **60**, 774-782.
147. Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, 1995.
148. Y. Benjamini and D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics*, 2001, **29**, 1165-1188.
149. A. Reiner, D. Yekutieli and Y. Benjamini, Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics (Oxford, England)*, 2003, **19**, 368-375.
150. Y. Benjamini and D. Yekutieli, Quantitative trait Loci analysis using the false discovery rate, *Genetics*, 2005, **171**, 783-790.
151. *SAS/STAT® 9.2 User's Guide.*, SAS Institute Inc., Cary, NC, USA, Cary, NC:, 2008.
152. J. Toedling, O. Skylar, T. Krueger, J. J. Fischer, S. Sperling, et al., Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts, *BMC bioinformatics*, 2007, **8**, 221.
153. R. Gottardo, Modeling and analysis of ChIP-chip experiments, *Methods in molecular biology (Clifton, N.J*, 2009, **567**, 133-143.
154. M. J. Buck, A. B. Nobel and J. D. Lieb, ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data, *Genome biology*, 2005, **6**, R97.
155. S. Keles, M. J. van der Laan, S. Dudoit and S. E. Cawley, Multiple testing methods for ChIP-Chip high density oligonucleotide array data, *J Comput Biol*, 2006, **13**, 579-613.
156. S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, et al., Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs, *Cell*, 2004, **116**, 499-509.
157. W. Li, C. A. Meyer and X. S. Liu, A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences, *Bioinformatics (Oxford, England)*, 2005, **21 Suppl 1**, i274-282.
158. H. Ji and W. H. Wong, TileMap: create chromosomal map of tiling array hybridizations, *Bioinformatics (Oxford, England)*, 2005, **21**, 3629-3636.

159. M. Zheng, L. O. Barrera, B. Ren and Y. N. Wu, ChIP-chip: data, model, and analysis, *Biometrics*, 2007, **63**, 787-796.
160. S. Keles, Mixture modeling for genome-wide localization of transcription factors, *Biometrics*, 2007, **63**, 10-21.
161. R. Gottardo, W. Li, W. E. Johnson and X. S. Liu, A flexible and powerful bayesian hierarchical model for ChIP-Chip experiments, *Biometrics*, 2008, **64**, 468-478.
162. W. Sun, M. J. Buck, M. Patel and I. J. Davis, Improved ChIP-chip analysis by a mixture model approach, *BMC bioinformatics*, 2009, **10**, 173.
163. D. Weese, A. K. Emde, T. Rausch, A. Doring and K. Reinert, RazerS--fast read mapping with sensitivity control, *Genome research*, 2009, **19**, 1646-1654.
164. B. Langmead, C. Trapnell, M. Pop and S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome biology*, 2009, **10**, R25.
165. R. Li, Y. Li, K. Kristiansen and J. Wang, SOAP: short oligonucleotide alignment program, *Bioinformatics (Oxford, England)*, 2008, **24**, 713-714.
166. J. C. Dohm, C. Lottaz, T. Borodina and H. Himmelbauer, Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, *Nucleic Acids Res*, 2008, **36**, e105.
167. O. Owolabi and D. R. McGregor, *Fast approximate string matching*, 1988.
168. P. Jokinen and E. Ukkonen, *Two algorithms for approximate string matching in static texts*, 1991.
169. R. Jothi, S. Cuddapah, A. Barski, K. Cui and K. Zhao, Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data, *Nucleic Acids Res*, 2008, **36**, 5221-5231.
170. C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, et al., A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, *Bioinformatics (Oxford, England)*, 2009, **25**, 1952-1958.
171. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, et al., Model-based analysis of ChIP-Seq (MACS), *Genome biology*, 2008, **9**, R137.
172. G. Tuteja, P. White, J. Schug and K. H. Kaestner, Extracting transcription factor targets from ChIP-Seq data, *Nucleic Acids Res*, 2009, **37**, e113.
173. T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Nature*, 2007, **448**, 553-560.
174. A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, et al., FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology, *Bioinformatics (Oxford, England)*, 2008, **24**, 1729-1730.
175. A. Alexa, J. Rahnenfuhrer and T. Lengauer, Improved scoring of functional groups from gene expression data by decorrelating GO graph structure, *Bioinformatics (Oxford, England)*, 2006, **22**, 1600-1607.
176. S. Falcon and R. Gentleman, Using GOstats to test gene lists for GO term association, *Bioinformatics (Oxford, England)*, 2007, **23**, 257-258.
177. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature genetics*, 2000, **25**, 25-29.
178. V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, et al., TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Res*, 2003, **31**, 374-378.
179. D. Vlieghe, A. Sandelin, P. J. De Bleser, K. Vleminckx, W. W. Wasserman, et al., A new generation of JASPAR, the open-access repository for transcription factor binding site profiles, *Nucleic Acids Res*, 2006, **34**, D95-97.

180. T. D. Schneider and R. M. Stephens, Sequence logos: a new way to display consensus sequences, *Nucleic Acids Res*, 1990, **18**, 6097-6100.
181. A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, et al., MATCH: A tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Res*, 2003, **31**, 3576-3579.
182. M. L. Bulyk, Computational prediction of transcription-factor binding site locations, *Genome biology*, 2003, **5**, 201.
183. H. G. Roider, A. Kanhere, T. Manke and M. Vingron, Predicting transcription factor affinities to DNA from a biophysical model, *Bioinformatics (Oxford, England)*, 2007, **23**, 134-141.
184. S. S. Zumdahl, *Chemical Principles*, 2002.
185. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, et al., The human genome browser at UCSC, *Genome research*, 2002, **12**, 996-1006.
186. J. Schlesinger, M. Schueler, M. Grunert, Fischer JJ, Zhang Q, et al., The Cardiac Transcription Network Modulated by Transcription Factors, Histone Modifications and MicroRNAs, *PLoS genetics*, 2010.
187. H. Klein and M. Vingron, Using transcription factor binding site co-occurrence to predict regulatory regions, *Genome informatics*, 2007, **18**, 109-118.
188. U. J. Pape, H. Klein and M. Vingron, Statistical detection of cooperative transcription factors with similarity adjustment, *Bioinformatics (Oxford, England)*, 2009, **25**, 2103-2109.
189. Z. Hu, B. Hu and J. F. Collins, Prediction of synergistic transcription factors by function conservation, *Genome biology*, 2007, **8**, R257.
190. P. Van Loo, S. Aerts, B. Thienpont, B. De Moor, Y. Moreau, et al., ModuleMiner - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues?, *Genome biology*, 2008, **9**, R66.
191. M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, et al., Assessing computational tools for the discovery of transcription factor binding sites, *Nature biotechnology*, 2005, **23**, 137-144.
192. B. R. Huber and M. L. Bulyk, Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data, *BMC bioinformatics*, 2006, **7**, 229.
193. X. Liu, D. L. Brutlag and J. S. Liu, BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes, *Pacific Symposium on Biocomputing*, 2001, 127-138.
194. J. D. Hughes, P. W. Estep, S. Tavazoie and G. M. Church, Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae, *Journal of molecular biology*, 2000, **296**, 1205-1214.
195. G. Pavesi, G. Mauri and G. Pesole, An algorithm for finding signals of unknown length in DNA sequences, *Bioinformatics (Oxford, England)*, 2001, **17 Suppl 1**, S207-214.
196. C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, et al., Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science (New York, N.Y*, 1993, **262**, 208-214.
197. D. Gusfield, *Algorithms on Strings, Trees and Sequences*, 1998.
198. T. L. Bailey and C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB*, 1994, **2**, 28-36.

199. A. M. McGuire, J. D. Hughes and G. M. Church, Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes, *Genome research*, 2000, **10**, 744-757.
200. Y. Pilpel, P. Sudarsanam and G. M. Church, Identifying regulatory networks by combinatorial analysis of promoter elements, *Nature genetics*, 2001, **29**, 153-159.
201. M. Friberg, P. von Rohr and G. Gonnet, Scoring functions for transcription factor binding site prediction, *BMC bioinformatics*, 2005, **6**, 84.
202. M. Owens, The Definitive Guide to SQLite, http://encompass.library.cornell.edu/cgi-bin/checkIP.cgi?access=gateway%5Fstandard%26url=http://dx.doi.org/10.1007/978-1-4302-0172-4
203. M. Koudritsky and E. Domany, Positional distribution of human transcription factor binding sites, *Nucleic acids research*, 2008, **36**, 6795-6805.
204. Y. Tabach, R. Brosh, Y. Buganim, A. Reiner, O. Zuk, et al., Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site, *PloS one*, 2007, **2**, e807.
205. F. J. Naya, B. L. Black, H. Wu, R. Bassel-Duby, J. A. Richardson, et al., Mitochondrial deficiency and cardiac sudden death in mice lacking the MEF2A transcription factor, *Nature medicine*, 2002, **8**, 1303-1309.
206. T. E. Royce, J. S. Rozowsky and M. B. Gerstein, Assessing the need for sequence-based normalization in tiling microarray experiments, *Bioinformatics (Oxford, England)*, 2007, **23**, 988-997.
207. J. W. Tukey, *Exploratory Data Analysis*, 1977.
208. R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, et al., Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Res*, 2003, **31**, e15.
209. G. K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, *Statistical applications in genetics and molecular biology*, 2004, **3**, Article3.
210. P. C. Scacheri, S. Davis, D. T. Odom, G. E. Crawford, S. Perkins, et al., Genome-wide analysis of menin binding provides insights into MEN1 tumorigenesis, *PLoS genetics*, 2006, **2**, e51.
211. A. Yang, Z. Zhu, P. Kapranov, F. McKeon, G. M. Church, et al., Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells, *Molecular cell*, 2006, **24**, 593-602.
212. S. R. Krig, V. X. Jin, M. C. Bieda, H. O'Geen, P. Yaswen, et al., Identification of genes directly regulated by the oncogene ZNF217 using chromatin immunoprecipitation (ChIP)-chip assays, *The Journal of biological chemistry*, 2007, **282**, 9703-9712.
213. R. Martone, G. Euskirchen, P. Bertone, S. Hartman, T. E. Royce, et al., Distribution of NF-kappaB-binding sites across human chromosome 22, *Proceedings of the National Academy of Sciences of the United States of America*, 2003, **100**, 12247-12252.
214. C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, et al., Transcriptional regulatory code of a eukaryotic genome, *Nature*, 2004, **431**, 99-104.
215. P. Phuc Le, J. R. Friedman, J. Schug, J. E. Brestelli, J. B. Parker, et al., Glucocorticoid receptor-dependent gene regulatory networks, *PLoS genetics*, 2005, **1**, e16.
216. Z. Hu, P. J. Killion and V. R. Iyer, Genetic reconstruction of a functional transcriptional regulatory network, *Nature genetics*, 2007, **39**, 683-687.
217. Y. S. Kwon, I. Garcia-Bassets, K. R. Hutt, C. S. Cheng, M. Jin, et al., Sensitive ChIP-DSL technology reveals an extensive estrogen receptor alpha-binding program on human gene promoters, *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**, 4852-4857.

218. M. Yu, L. Riva, H. Xie, Y. Schindler, T. B. Moran, et al., Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis, *Molecular cell*, 2009, **36**, 682-695.
219. A. A. Alekseyenko, S. Peng, E. Larschan, A. A. Gorchakov, O. K. Lee, et al., A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome, *Cell*, 2008, **134**, 599-609.
220. H. Choi, A. I. Nesvizhskii, D. Ghosh and Z. S. Qin, Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data, *Bioinformatics (Oxford, England)*, 2009, **25**, 1715-1721.
221. P. Qiu and L. Li, Histone acetylation and recruitment of serum responsive factor and CREB-binding protein onto SM22 promoter during SM22 gene expression, *Circulation research*, 2002, **90**, 858-865.
222. Z. Wang, C. Zang, K. Cui, D. E. Schones, A. Barski, et al., Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes, *Cell*, 2009, **138**, 1019-1031.
223. C. M. Koch, R. M. Andrews, P. Flicek, S. C. Dillon, U. Karaoz, et al., The landscape of histone modifications across 1% of the human genome in five human cell lines, *Genome research*, 2007, **17**, 691-707.
224. N. Vo and R. H. Goodman, CREB-binding protein and p300 in transcriptional regulation, *The Journal of biological chemistry*, 2001, **276**, 13505-13508.
225. N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, et al., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome, *Nature genetics*, 2007, **39**, 311-318.
226. A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, et al., ChIP-seq accurately predicts tissue-specific activity of enhancers, *Nature*, 2009, **457**, 854-858.
227. M. G. Pray-Grant, J. A. Daniel, D. Schieltz, J. R. Yates, 3rd and P. A. Grant, Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation, *Nature*, 2005, **433**, 434-438.
228. J. Wysocka, T. Swigut, T. A. Milne, Y. Dou, X. Zhang, et al., WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development, *Cell*, 2005, **121**, 859-872.
229. D. G. Martin, D. E. Grimes, K. Baetz and L. Howe, Methylation of histone H3 mediates the association of the NuA3 histone acetyltransferase with chromatin, *Molecular and cellular biology*, 2006, **26**, 3018-3028.
230. O. G. McDonald and G. K. Owens, Programming smooth muscle plasticity with chromatin dynamics, *Circulation research*, 2007, **100**, 1428-1441.
231. O. G. McDonald, B. R. Wamhoff, M. H. Hoofnagle and G. K. Owens, Control of SRF binding to CArG box chromatin regulates smooth muscle gene expression in vivo, *The Journal of clinical investigation*, 2006, **116**, 36-48.
232. W. N. Venables, B. Ripley and W. N. Venables, *Modern applied statistics with S*, Springer, 2002.
233. P. Ellinghaus, R. J. Scheubel, D. Dobrev, U. Ravens, J. Holtz, et al., Comparing the global mRNA expression profile of human atrial and ventricular myocardium with high-density oligonucleotide arrays, *The Journal of thoracic and cardiovascular surgery*, 2005, **129**, 1383-1390.
234. R. Tabibiazar, R. A. Wagner, A. Liao and T. Quertermous, Transcriptional profiling of the heart reveals chamber-specific gene expression patterns, *Circulation research*, 2003, **93**, 1193-1201.
235. K. Fellenberg, N. C. Hauser, B. Brors, A. Neutzner, J. D. Hoheisel, et al., Correspondence analysis applied to microarray data, *Proceedings of the National Academy of Sciences of the United States of America*, 2001, **98**, 10781-10786.

236. P. J. Huber, *Robust statistics*, John Wiley & Sons, New York [etc.], 1981.
237. L. A. Goff, J. Davila, R. Jornsten, S. Keles and R. P. Hart, Bioinformatic analysis of neural stem cell differentiation, *J Biomol Tech*, 2007, **18**, 205-212.
238. S. Y. Kim and Y. Kim, Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data, *BMC bioinformatics*, 2006, **7**, 330.
239. S. Nelander, E. Larsson, E. Kristiansson, R. Mansson, O. Nerman, et al., Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals, *BMC genomics*, 2005, **6**, 68.
240. V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, et al., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic acids research*, 2006, **34**, D108-110.
241. U. Gerland, J. D. Moroz and T. Hwa, Physical constraints and functional characteristics of transcription factor-DNA interaction, *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**, 12015-12020.
242. R. R. Copley, M. Totrov, J. Linnell, S. Field, J. Ragoussis, et al., Functional conservation of Rel binding sites in drosophilid genomes, *Genome research*, 2007, **17**, 1327-1335.
243. I. S. Skerjanc, H. Petropoulos, A. G. Ridgeway and S. Wilton, Myocyte enhancer factor 2C and Nkx2-5 up-regulate each other's expression and initiate cardiomyogenesis in P19 cells, *The Journal of biological chemistry*, 1998, **273**, 34904-34910.
244. M. Tanaka, Z. Chen, S. Bartunkova, N. Yamasaki and S. Izumo, The cardiac homeobox gene Csx/Nkx2.5 lies genetically upstream of multiple genes essential for heart development, *Development (Cambridge, England)*, 1999, **126**, 1269-1280.
245. D. G. McFadden, J. Charite, J. A. Richardson, D. Srivastava, A. B. Firulli, et al., A GATA-dependent right ventricular enhancer controls dHAND transcription in the developing heart, *Development (Cambridge, England)*, 2000, **127**, 5331-5341.
246. E. Dodou, M. P. Verzi, J. P. Anderson, S. M. Xu and B. L. Black, Mef2c is a direct transcriptional target of ISL1 and GATA factors in the anterior heart field during mouse embryonic development, *Development (Cambridge, England)*, 2004, **131**, 3931-3942.
247. A. Rojas, S. De Val, A. B. Heidt, S. M. Xu, J. Bristow, et al., Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by Forkhead and GATA transcription factors through a distal enhancer element, *Development (Cambridge, England)*, 2005, **132**, 3405-3417.
248. W. Nishida, M. Nakamura, S. Mori, M. Takahashi, Y. Ohkawa, et al., A triad of serum response factor and the GATA and NK families governs the transcription of smooth and cardiac muscle genes, *The Journal of biological chemistry*, 2002, **277**, 7308-7317.
249. R. Jelier, M. J. Schuemie, A. Veldhoven, L. C. Dorssers, G. Jenster, et al., Anni 2.0: a multipurpose text-mining tool for the life sciences, *Genome biology*, 2008, **9**, R96.
250. B. Snel, G. Lehmann, P. Bork and M. A. Huynen, STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene, *Nucleic acids research*, 2000, **28**, 3442-3444.
251. T. Barrett and R. Edgar, Gene expression omnibus: microarray data storage, submission, retrieval, and analysis, *Methods in enzymology*, 2006, **411**, 352-369.
252. H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, et al., ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression, *Nucleic acids research*, 2009, **37**, D868-872.
253. K. P. Yee, D. Fisher, R. Dhamija and Hearst, *M.: Animated exploration of dynamic graphs with radial layout*.

254. U. Brandes, M. Eiglsperger, I. Herman, M. Himsolt and M. Marshall, *GraphML progress report: Structural layer proposal*, 2002.
255. J. Heer, S. K. Card and J. Landay, *A.: prefuse: a toolkit for interactive information visualization*.
256. C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, et al., STRING: a database of predicted functional associations between proteins, *Nucleic acids research*, 2003, **31**, 258-261.
257. C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, et al., STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucleic acids research*, 2005, **33**, D433-437.
258. C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, et al., STRING 7--recent developments in the integration and prediction of protein interactions, *Nucleic acids research*, 2007, **35**, D358-362.
259. L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, et al., STRING 8--a global view on proteins and their functional interactions in 630 organisms, *Nucleic acids research*, 2009, **37**, D412-416.
260. R. Jelier, G. Jenster, L. C. Dorssers, B. J. Wouters, P. J. Hendriksen, et al., Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation, *BMC bioinformatics*, 2007, **8**, 14.
261. P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, et al., TXTGate: profiling gene groups with text-based information, *Genome biology*, 2004, **5**, R43.
262. R. Jelier, M. J. Schuemie, P. J. Roes, E. M. van Mulligen and J. A. Kors, Literature-based concept profiles for gene annotation: the issue of weighting, *International journal of medical informatics*, 2008, **77**, 354-362.
263. F. A. Stennard, M. W. Costa, D. A. Elliott, S. Rankin, S. J. Haast, et al., Cardiac T-box factor Tbx20 directly interacts with Nkx2-5, GATA4, and GATA5 in regulation of gene expression in the developing heart, *Developmental biology*, 2003, **262**, 206-224.
264. E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, et al., Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, 2007, **447**, 799-816.
265. P. J. Farnham, Insights from genomic profiling of transcription factors, *Nature reviews*, 2009, **10**, 605-616.
266. J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann and N. M. Luscombe, A census of human transcription factors: function, expression and evolution, *Nature reviews*, 2009, **10**, 252-263.
267. A. J. Ruthenburg, H. Li, D. J. Patel and C. D. Allis, Multivalent engagement of chromatin modifications by linked binding modules, *Nat Rev Mol Cell Biol*, 2007, **8**, 983-994.
268. S. Dalton, R. Marais, J. Wynne and R. Treisman, Isolation and characterization of SRF accessory proteins, *Philosophical transactions of the Royal Society of London*, 1993, **340**, 325-332.
269. O. W. Prall, M. K. Menon, M. J. Solloway, Y. Watanabe, S. Zaffran, et al., An Nkx2-5/Bmp2/Smad1 negative feedback loop controls heart progenitor specification and proliferation, *Cell*, 2007, **128**, 947-959.
270. C. Tuerk and L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, *Science (New York, N.Y*, 1990, **249**, 505-510.

271. R. Stoltenburg, C. Reinemann and B. Strehlitz, SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands, *Biomolecular engineering*, 2007, **24**, 381-403.
272. H. Akazawa and I. Komuro, Cardiac transcription factor Csx/Nkx2-5: Its role in cardiac development and diseases, *Pharmacology & therapeutics*, 2005, **107**, 252-268.
273. K. L. Clark, K. E. Yutzey and D. W. Benson, Transcription factors and congenital heart defects, *Annual review of physiology*, 2006, **68**, 97-121.
274. A. L'Honore, V. Rana, N. Arsic, C. Franckhauser, N. J. Lamb, et al., Identification of a new hybrid serum response factor and myocyte enhancer factor 2-binding element in MyoD enhancer required for MyoD expression during myogenesis, *Molecular biology of the cell*, 2007, **18**, 1992-2001.
275. G. K. Smyth, *Limma: linear models for microarray data*, 2005.
276. Y. H. Loh, Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, et al., The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells, *Nature genetics*, 2006, **38**, 431-440.
277. T. J. Page, D. Sikder, L. Yang, L. Pluta, R. D. Wolfinger, et al., Genome-wide analysis of human HSF1 signaling reveals a transcriptional program linked to cellular adaptation and survival, *Molecular bioSystems*, 2006, **2**, 627-639.
278. T. Palomero, D. T. Odom, J. O'Neil, A. A. Ferrando, A. Margolin, et al., Transcriptional regulatory networks downstream of TAL1/SCL in T-cell acute lymphoblastic leukemia, *Blood*, 2006, **108**, 986-992.
279. F. Gao, B. C. Foat and H. J. Bussemaker, Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data, *BMC bioinformatics*, 2004, **5**, 31.
280. M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch and R. A. Young, A chromatin landmark and transcription initiation at most promoters in human cells, *Cell*, 2007, **130**, 77-88.
281. A. Gitter, Z. Siegfried, M. Klutstein, O. Fornes, B. Oliva, et al., Backup in gene regulatory networks explains differences between binding and knockout results, *Molecular systems biology*, 2009, **5**, 276.
282. Q. Lin, J. Schwarz, C. Bucana and E. N. Olson, Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C, *Science (New York, N.Y*, 1997, **276**, 1404-1407.
283. L. H. Vong, M. J. Ragusa and J. J. Schwarz, Generation of conditional Mef2cloxP/loxP mice for temporal- and tissue-specific analyses, *Genesis*, 2005, **43**, 43-48.
284. R. H. Morse, Epigenetic marks identify functional elements, *Nature genetics*, 2010, **42**, 282-284.
285. T. A. McKinsey, C. L. Zhang and E. N. Olson, Control of muscle development by dueling HATs and HDACs, *Current opinion in genetics & development*, 2001, **11**, 497-504.
286. J. Backs and E. N. Olson, Control of cardiac growth by histone acetylation/deacetylation, *Circulation research*, 2006, **98**, 15-24.
287. H. Wang, R. Cao, L. Xia, H. Erdjument-Bromage, C. Borchers, et al., Purification and functional characterization of a histone H3-lysine 4-specific methyltransferase, *Molecular cell*, 2001, **8**, 1207-1217.
288. M. Haberland, R. L. Montgomery and E. N. Olson, The many roles of histone deacetylases in development and physiology: implications for disease and therapy, *Nature reviews*, 2009, **10**, 32-42.

289. S. Sperling, C. H. Grimm, I. Dunkel, S. Mebus, H. P. Sperling, et al., Identification and functional analysis of CITED2 mutations in patients with congenital heart defects, *Human mutation*, 2005, **26**, 575-582.
290. S. T. MacDonald, S. D. Bamforth, C. M. Chen, C. R. Farthing, A. Franklyn, et al., Epiblastic Cited2 deficiency results in cardiac phenotypic heterogeneity and provides a mechanism for haploinsufficiency, *Cardiovascular research*, 2008, **79**, 448-457.
291. F. Bailliard and R. H. Anderson, Tetralogy of Fallot, *Orphanet journal of rare diseases*, 2009, **4**, 2.
292. S. Hammer, M. Toenjes, M. Lange, J. J. Fischer, I. Dunkel, et al., Characterization of TBX20 in human hearts and its regulation by TFAP2, *Journal of cellular biochemistry*, 2008, **104**, 1022-1033.
293. J. K. Takeuchi, M. Mileikovskaia, K. Koshiba-Takeuchi, A. B. Heidt, A. D. Mori, et al., Tbx20 dose-dependently regulates transcription factor networks required for mouse heart and motoneuron development, *Development (Cambridge, England)*, 2005, **132**, 2463-2474.
294. E. P. Kirk, M. Sunde, M. W. Costa, S. A. Rankin, O. Wolstein, et al., Mutations in cardiac T-box factor gene TBX20 are associated with diverse cardiac pathologies, including defects of septation and valvulogenesis and cardiomyopathy, *American journal of human genetics*, 2007, **81**, 280-291.
295. T. F. Plageman, Jr. and K. E. Yutzey, Differential expression and function of Tbx5 and Tbx20 in cardiac development, *The Journal of biological chemistry*, 2004, **279**, 19026-19034.
296. G. Nemer, F. Fadlalah, J. Usta, M. Nemer, G. Dbaibo, et al., A novel mutation in the GATA4 gene in patients with Tetralogy of Fallot, *Human mutation*, 2006, **27**, 293-294.
297. T. Park, S. G. Yi, S. H. Kang, S. Lee, Y. S. Lee, et al., Evaluation of normalization methods for microarray data, *BMC bioinformatics*, 2003, **4**, 33.
298. W. Wu, E. P. Xing, C. Myers, I. S. Mian and M. J. Bissell, Evaluation of normalization methods for cDNA microarray data by k-NN classification, *BMC bioinformatics*, 2005, **6**, 191.
299. W. E. Johnson, W. Li, C. A. Meyer, R. Gottardo, J. S. Carroll, et al., Model-based analysis of tiling-arrays for ChIP-chip, *Proceedings of the National Academy of Sciences of the United States of America*, 2006, **103**, 12457-12462.
300. J. S. Song, W. E. Johnson, X. Zhu, X. Zhang, W. Li, et al., Model-based analysis of two-color arrays (MA2C), *Genome biology*, 2007, **8**, R178.
301. H. R. Chung, D. Kostka and M. Vingron, A physical model for tiling array analysis, *Bioinformatics (Oxford, England)*, 2007, **23**, i80-86.
302. T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, et al., A high-resolution map of active promoters in the human genome, *Nature*, 2005, **436**, 876-880.
303. Y. Qi, A. Rolfe, K. D. MacIsaac, G. K. Gerber, D. Pokholok, et al., High-resolution computational models of genome binding events, *Nature biotechnology*, 2006, **24**, 963-970.
304. D. J. Reiss, M. T. Facciotti and N. S. Baliga, Model-based deconvolution of genome-wide DNA binding, *Bioinformatics (Oxford, England)*, 2008, **24**, 396-403.
305. E. G. Wilbanks and M. T. Facciotti, Evaluation of algorithm performance in ChIP-seq peak detection, *PloS one*, **5**, e11471.
306. S. Rahmann, T. Muller and M. Vingron, On the power of profiles for transcription factor binding site detection, *Statistical applications in genetics and molecular biology*, 2003, **2**, Article7.

307. A. R. Borneman, T. A. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, et al., Divergence of transcription factor binding sites across related yeast species, *Science (New York, N.Y*, 2007, **317**, 815-819.
308. R. K. Bradley, X. Y. Li, C. Trapnell, S. Davidson, L. Pachter, et al., Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species, *PLoS biology*, **8**, e1000343.
309. D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, et al., Tissue-specific transcriptional regulation has diverged significantly between human and mouse, *Nature genetics*, 2007, **39**, 730-732.
310. N. Friedman, M. Linial, I. Nachman and D. Pe'er, Using Bayesian networks to analyze expression data, *J Comput Biol*, 2000, **7**, 601-620.
311. J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink and E. D. Jarvis, Advances to Bayesian network inference for generating causal networks from observational biological data, *Bioinformatics (Oxford, England)*, 2004, **20**, 3594-3603.
312. A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, et al., A systematic comparison and evaluation of biclustering methods for gene expression data, *Bioinformatics (Oxford, England)*, 2006, **22**, 1122-1129.
313. J. Cui, J. Liu, Y. Li and T. Shi, Integrative Identification of Arabidopsis Mitochondrial Proteome and Its Function Exploitation through Protein Interaction Network, *PloS one*, 2011, **6**, e16022.
314. M. M. Elhefnawi, A. A. Youssif, A. Z. Ghalwash and W. H. Behaidy, An integrated methodology for mining promiscuous proteins: a case study of an integrative bioinformatics approach for hepatitis C virus non-structural 5A protein, *Advances in experimental medicine and biology*, 2010, **680**, 299-305.
315. R. Konig, S. Stertz, Y. Zhou, A. Inoue, H. H. Hoffmann, et al., Human host factors required for influenza virus replication, *Nature*, 2010, **463**, 813-817.
316. P. Arrigo, P. P. Cardo and A. Izzotti, Proteomic-based screening of miRNAs metabolic pathway targeting, *International journal of computational biology and drug design*, 2010, **3**, 164-173.
317. G. Pandey, B. Zhang, A. N. Chang, C. L. Myers, J. Zhu, et al., An integrative multi-network and multi-classifier approach to predict genetic interactions, *PLoS computational biology*, **6**.
318. C. Kaleta, A. Gohler, S. Schuster, K. Jahreis, R. Guthke, et al., Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis, *BMC systems biology*, 2010, **4**, 116.
319. V. Kaimal, D. Sardana, E. E. Bardes, R. C. Gudivada, J. Chen, et al., Integrative systems biology approaches to identify and prioritize disease and drug candidate genes, *Methods in molecular biology (Clifton, N.J*, 2011, **700**, 241-259.
320. S. Iadevaia, Y. Lu, F. C. Morales, G. B. Mills and P. T. Ram, Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis, *Cancer research*, 2010, **70**, 6704-6714.
321. S. Sperling, Systems biology approaches to heart development and congenital heart disease, *In preparation*, 2011.
322. J. B. Huang, Y. L. Liu, P. W. Sun, X. D. Lv, M. Du, et al., Molecular mechanisms of congenital heart disease, *Cardiovasc Pathol*, **19**, e183-193.
323. I. Mateo Leach, P. van der Harst and R. A. de Boer, Pharmacoepigenetics in heart failure, *Current heart failure reports*, **7**, 83-90.
324. S. A. Miller and A. S. Weinmann, An essential interaction between T-box proteins and histone-modifying enzymes, *Epigenetics*, 2009, **4**, 85-88.
325. R. Papait and G. Condorelli, Epigenetics in heart failure, *Annals of the New York Academy of Sciences*, **1188**, 159-164.

326. L. Cartegni, S. L. Chew and A. R. Krainer, Listening to silence and understanding nonsense: exonic mutations that affect splicing, *Nature reviews*, 2002, **3**, 285-298.
327. M. Krawczak, J. Reiss and D. N. Cooper, The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences, *Human genetics*, 1992, **90**, 41-54.
328. R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, et al., Regulation of alternative splicing by histone modifications, *Science (New York, N.Y*, 2010, **327**, 996-1000.
329. T. Thum, P. Galuppo, C. Wolf, J. Fiedler, S. Kneitz, et al., MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure, *Circulation*, 2007, **116**, 258-267.
330. T. Thum, D. Catalucci and J. Bauersachs, MicroRNAs: novel regulators in cardiac development and disease, *Cardiovascular research*, 2008, **79**, 562-570.
331. T. Thum, C. Gross, J. Fiedler, T. Fischer, S. Kissler, et al., MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts, *Nature*, 2008, **456**, 980-984.
332. J. Krutzfeldt, N. Rajewsky, R. Braich, K. G. Rajeev, T. Tuschl, et al., Silencing of microRNAs in vivo with 'antagomirs', *Nature*, 2005, **438**, 685-689.

6. Abbreviations

ANOVA	Analysis of variance
AoArch	Aortic arch
ASDII	Atrial septal defect, secundum type
AVSD	Atrioventricular septal defect
Bpsys	Systolic blood pressure
CHD	Congenital heart disease
ChIP	Chromatin immunoprecipitation
ChIP-chip	ChIP followed by array analysis
ChIP-seq	ChIP followed by next-generation sequencing
DORV	Double outlet right ventricle
E18.5	Embryonic state 18 ½ days after fertilization
FDR	False discovery rate
GO	Gene ontology
H3ac	Acetylation of lysine 9 and 14 on histone 3
H3K4	Lysin 4 of histone 3
H3K4me	Mono-methylation of lysine 4 on histone 3
H3K4me2	Di-methylation of lysine 4 on histone 3
H3K4me3	Tri-methylation of lysine 4 on histone 3
H4ac	Acetylation of lysine 5, 8,12 and 16 on histone 4
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HLHS	Hypoplastic leaft heart syndrome
IAS	Interatrial septal defect
Infund	Infundibular
Insuff	Insufficiency
IVS	Interventricular septum
LA	Left atrium
LR	Left to right
LSVC	Left superior caval vein present
LV	Left ventricle
miRNA	MicroRNA
miRNA-seq	Next-generation sequencing of miRNAs
MPA	Main pulmonary artery
P0.5	Postnatal stage ½ days after birth
P4.5	Postnatal stage 4 ½ days after birth
PCR	Polymerase chain reaction
PDA	Persistent ductus arteriosus
Perim	Perimembranous
PFO	Patent Foramen ovale

PHD	Plant homeodomain
PS	Pulmonary stenosis
PV	Pulmonary valve
PWM	Position weight matrix
qPCR	Quantitative real-time PCR
Q-Q plot	Quantile-quantile plot
RA	Right atrium
RISC	RNA-induced silencer complex
RL	Right to left
RNAi	RNA interference
RV	Right ventricle
siRNA	Short interference RNA
Sysgrad	Systolic gradient
TF	Transcription factor
TFBS	Transcription factor binding site
TGA	Transposition of the great arteries
TOF	Tetralogy of Fallot
TSS	Transcriptional start site
UTR	Untranslated region
VSD	Ventricular septal defect

7. Summary

This work presents the bioinformatic part of an integrative approach to analyze transcription networks taking the vertebrate heart as a model. Bioinformatic analyses of a number of high-throughput experiments elucidated regulatory dependencies that drive correct spatial and temporal transcription by the interplay between combinatorial transcription factors binding and co-occurring histone modifications. Furthermore, key genes were characterized in terms of their association with cardiac disease and a database to store and visualize current knowledge in heart and muscle gene regulatory associations was implemented.

The analysis of the four key cardiac transcription factors Gata4, Mef2a, Nkx2.5 and Srf using binding site detection by chromatin immunoprecipitation followed by array analysis (ChIP-chip) and knockdown of the respective factors using siRNA experiments revealed a high overlap in terms of their binding sites as well as their regulated target genes. Interestingly, those genes that had a high number of binding factors were less likely to be differentially expressed in the knockdown of the four TFs. This finding points to a buffering effect based on combinatorial binding, in which the remaining factors stabilize the gene expression even if a single factor is missing. Co-occurrence analysis with four activating histone modification revealed that the activating potential of Gata4, Nkx2.5 and Srf was highly dependent on the co-occurrence with histone modification marks, while no such dependency could be revealed for Mef2a. For Srf, the number of possible interacting histone modifications was narrowed down to histone 3 acetylation as the only important factor. This result was further substantiated by additional chromatin immunoprecipitation followed by next-generation Illumina sequencing (ChIP-seq), which also revealed that the presence of histone 3 acetylation tags had a buffering effect on the expression of Srf targets even after knockdown of this TF.

To gain further insights into Srf-driven gene transcription, the binding of Srf, the acetyltransferase p300 and the presence of histone 3 acetylation and histone 3 lysine 4 di-methylation were subsequently analyzed in selected regulatory regions in a time-series of mouse hearts around birth using chromatin immunoprecipitation followed by real-time qPCR measurements. The analysis revealed a high correlation in the enrichments of the individual factors over time. One- and two-factor linear models confirmed this correlation and substantiated a link between Srf and histone 3 acetylation through p300. However, they also implied so far unknown ways of coupling between these two regulatory factors. Changes in the analyzed factors were shown to have a regulatory input on the expression levels of nearby genes.

To investigate the significance of the found regulatory implications, a selected set of 42 genes were screened in a cohort of patients with a broad panel of congenital heart disease. Using linear models, specific molecular portraits associated to phenotypic subgroups could be identified. Further, using correlated expression and prediction of transcription factor binding sites, which was optimized using the previously analyzed ChIP datasets, cardiac regulatory networks could be revealed, which were verified using a combination of experimental, literature and bioinformatic datasets.

Finally, the ‘CArdiovascular Regulatory INteraction’ database was implemented, which integrates experimental results from several species as well as publicly available annotations and offers a sophisticated user interface which provides an easy and comfortable data overview using dynamic network representations and detailed information for individual genes at the same time. The database

was developed to enable the integrated view of data generated within the European project HeartRepair and to promote future research in this field.

In summary, the presented analyses revealed high complexity of the genetic and epigenetic levels of cardiac gene regulatory networks and a high interdependency between these.

8. Zusammenfassung

Die vorliegende Arbeit stellt den bioinformatischen Teil einer integrativen Analyse transkriptioneller Netzwerke am Modell des Wirbeltierherzens vor. Mit Hilfe von Datensätzen gewonnen durch aktuellste experimentelle Techniken wurden dabei Abhängigkeiten innerhalb der regulierenden Netzwerke aufgezeigt. Es wurde das Zusammenspiel von gemeinsamer Bindung von Transkriptionsfaktoren (TF) an Genpromotoren und das gleichzeitige Vorkommen von Histonmodifikationen näher untersucht. Weitere Schwerpunkt waren die Bestimmung von deregulierten Genen in Patienten mit angeborenen Herzfehlern und der Aufbau einer Datenbank zur Speicherung und Visualisierung des gegenwärtig vorhandenen Wissens über herz- und muskelspezifische Gene und deren Regulation.

Die Analyse von Chromatin-Immunoprezipitation mit anschließender Mikro-Array-Detektion für die Transkriptionsfaktoren Gata4, Mef2a, Nkx2.5 und Srf sowie die Datenanalyse von siRNA-vermittelten Knockdowns derselben TF ergab eine hohen Zahl an gemeinsamen Bindestellen und regulierten Zielgenen. Interessanterweise waren genau die Gene, welche die höchste Zahl an Bindestellen für die untersuchten TF aufwiesen, am seltensten im Knockdown der jeweiligen Transkriptionsfaktoren differentiell exprimiert. Dies weist auf einen potentiellen Puffereffekt durch kombinatorische Bindung mehrerer TF hin. Dabei können Gata4, Mef2a, Nkx2.5 und Srf die Expression von Zielgenen auch dann aufrecht erhalten, wenn ein einzelner TF nicht mehr vorhanden ist. Die Untersuchung des gemeinsamen Auftretens von TF-Bindestellen und Histonmodifikationen verdeutlichte, dass Gata4, Nkx2.5 und Srf ihr aktivierendes Potential nur im Zusammenspiel mit Histonmodifikationen entfalten können. Ein vergleichbarer Zusammenhang für Mef2a wurde nicht gefunden. Für Srf konnte die Acetylierung von Histone 3 (H3) als einzigen bestimmenden Faktor isolieren werden. Weitergehende Chromatin-Immunopräzipitations-Experimente mit anschließender Next-Generation-Sequenzierung (ChIP-seq) untermauerten die Ergebnisse und deuteten zudem einen Puffereffekt der H3-Acetylierung auf die Expression von Srf-Zielgenen in dessen Knockdown an.

In einer auf diese Ergebnisse aufbauenden Analyse wurde der Zusammenhang zwischen der Bindung von Srf und der Acetyltransferase p300 sowie den zwei Histonmodifikationen H3-Acetylierung und H3-Lysin-4-Dimethylierung näher untersucht. Dazu wurden potentiell regulierende DNA-Regionen in einer Zeitreihenanalyse von Mauseherzen vor und nach der Geburt durch Chromatin-Immunopräzipitation gefolgt von quantitativer Real-Time-PCR untersucht. Die Analyse ergab eine hohe zeitliche Korrelation der untersuchten Faktoren, die durch den Einsatz linearer Modelle weiter untermauert werden konnte. Dabei konnte der bereits vorher bekannte funktionelle Link zwischen p300, Srf und H3-Acetylierung bestätigt werden. Zudem ergab die Analyse, dass es potentiell weitere von p300 unabhängige regulierende Mechanismen der H3-Acetylierung durch Srf gibt. Schließlich konnte gezeigt werden, dass eine Veränderung in der Bindung der untersuchten TF und Histonmodifikationen sich auf die Expression benachbarter Gene auswirkt.

Nachdem die vorangegangenen Untersuchungen sich auf transkriptionelle Netzwerke in Zellkultur und Mausherzen fokussierte, wurde abschließend die Expression von 42 ausgewählten Genen in Patienten mit verschiedensten angeborenen Herzfehlern untersucht. Dabei konnten phänotypischen Subgruppen spezifische molekulare Expressionsmuster zugeordnet werden. Mit Hilfe einer Kombination aus korrelierten Expressionprofilen und der Vorhersage von Transkriptionsfaktorbindestellen, die zuvor anhand der bereits untersuchten ChIP-Datensätze optimiert worden war, konnten daraufhin

transkriptionelle kardiale Regulationsnetzwerke vorhergesagt werden. Diese wurden mit Hilfe von mehreren biochemischen, bioinformatischen und Literaturdatensätzen verifiziert.

Als letzter Teil der Studie wurde die „CArdiovascular Regulatory INteraction“-Datenbank implementiert. Diese integriert aktuelle experimentelle Ergebnisse aus verschiedenen Organismen mit öffentlich verfügbaren Annotationsdatensätzen. Besonderer Wert wurde hierbei auf die Realisierung eines fortgeschrittenen Visualisierungsmoduls gelegt, das sowohl gespeicherte Netzwerke in dynamischer Form darstellen kann als auch Informationen für einzelne Gene bereitstellt. Die Datenbank soll eine Grundlage schaffen für eine weiterführende Analyse der erarbeiteten Ergebnisse sowohl von bioinformatischer als auch von biologischer Seite.

Zusammengefaßt ergaben die durchgeführten Analysen ein hohes Maß an Komplexität innerhalb der genetischen und epigenetischen Ebenen der Genregulation mit vielfachen Verknüpfungen.

9. Individual Contributions

All analyses described in this study have been performed in the group of Prof. Dr. Silke R. Sperling. The experiments investigated have been performed by biochemists from the Sperling group, including (in alphabetical order) Ilona Dunkel, Jenny J. Fischer, Stefanie Hammer, Martin Lange, Jenny Schlesinger, Martje Tönjes and Qin Zhang. To clarify my part in the data analyses I will in the following list my individual contributions in detail.

For the analysis described in section 3.1 of the results chapter, I performed the ChIP-chip peak calling, the analysis of found TFBS including the motif discovery, the co-occurrence of TFBS, the analysis of the siRNA experiments and the overlap between ChIP-chip and siRNA results together with Jörn Tödling and Tammo Krüger. The ChIP-seq mapping and peak calling were performed by Marcel Grunert. Together with him I analyzed the overlap between ChIP-chip and ChIP-seq TFBS. I analyzed the influence of histone modifications and Srf marks on gene expression.

For the study of the qPCR time-series data (section 3.2) I selected the regulatory regions and performed all of the bioinformatic analysis.

For the study of patient material (section 3.3) I performed all of the bioinformatic analysis, with the exceptions of the initial phenotype analysis, the expression data normalization and the correspondence analysis which have been performed by Utz. J. Pape.

The CARIN database and it's user interface (section 3.4) were solely implemented by me. Marcel Grunert participated in the conception and discussion of the user interface.

All analysis presented in this thesis have been discussed and supervised by Prof. Dr. Silke R. Sperling.

Curriculum Vitae

For reasons of data protection, the curriculum vitae is not included in the online version.

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

List of Own Publications

Publications

Schueler M*, Zhang Q*, Schlesinger J, Tönjes M, Sperling SR. 2011.„Dynamics of histone modifications and transcription factor binding during cardiac maturation in mice". In preparation. *Equal contribution.

Schlesinger J*, **Schueler M***, Grunert M*, Fischer JJ*, Zhang Q, Krueger T, Lange M, Tönjes M, Dunkel I, Sperling SR. 2011. „The Cardiac Transcription Network Modulated by Gata4, Mef2a, Nkx2.5 and Srf, Histone Modifications and MicroRNAs".*PLoS Genet* 7(2). * Equal contribution.

Schlesinger J, Tönjes M, **Schueler M**, Zhang Q, Dunkel I, Sperling SR. 2010. „Evaluation of the LightCycler© 1536 Instrument for high-throughput quantitative real-time PCR". *Methods* 50(4): 19-22.

Toenjes M*, **Schueler M***, Hammer S, Pape UJ, Berger F, Vingron M, Sperling S. 2008. "Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes". *Molecular BioSystems* 4(6): 589-598. * Equal contribution.

Fischer JJ, Toedling J, Krueger T, **Schueler M**, Huber W, Sperling S. 2008. „Combinatorial effects of four histone modifications in transcription and differentiation" *Genomics* 91/1: 41-51.

Purmann A, Toedling J, **Schueler M**, Carninci P, Lehrach H, Hayashizaki Y, Huber W, Sperling S. 2007. "Genomic organization of transcriptomes in mammals: Co-regulation and co-functionality" *Genomics* 89: 580-587.

Koch I, **Schueler M,** Heiner M. 2005. "STEPP-Search Tool for Exploration of Petri net Paths: a new tool for Petri net-based path analysis in biochemical networks". *In Silico Biology* 5(2):129-37.

Conference Talks

Title: "CARIN – The CArdiovascular Regulatory INteraction database". *HeartRepair Annual Conference*, April 5-8, 2009, Berlin, Germany

Title: "CARIN - A web database integrating gene expression with transcription and epigenetic regulation in cardiomyocytes". *HeartRepair Annual Conference*, April 19-21, 2008, Madrid, Spain.

Selected Poster Presentations

Schueler M*, Grunert M*, Schlesinger J*, Fischer JJ, Zhang Q, Dunkel I, Sperling SR. „Analyzing the Cardiac Transcription Network Driven by the Interplay of Transcription Factors, Histone Modifications and MicroRNAs". *RECOMB*, August 12-15, 2010, Lisbon, Portugal. *Equal contribution.

Selected Poster Presentations continued

Schueler M*, Tönjes M*, Hammer S, Pape UJ, Berger F, Vingron M, Sperling SR. "Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes". *ISMB/ECCB*, June 25 – July 2, 2009, Stockholm, Sweden. *Equal contribution.

Schueler M, Tönjes M, Fischer JJ, Hammer S, Pape UJ, Vingron M, Sperling SR. „Interactive framework for the creation, visualization and analysis of regulatory networks". *XX International Congress of Genetics*, July 12-17, 2008, Berlin, Germany.

Schueler M, Tönjes M, Fischer JJ, Hammer S, Pape UJ, Vingron M, Sperling SR. „Interactive framework for the creation, visualization and analysis of regulatory networks". *GCB*, September 26-28, 2007, Potsdam, Germany.

Schueler M, Koch I. "STEPP - Search Tool for Exploration of Petri net Paths: a new tool for Petri net-based path analysis in biochemical networks". *ICSB*, October 9-13, 2004, Heidelberg, Germany.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel in Anspruch genommen habe. Ich versichere, dass diese Arbeit in dieser oder anderer Form keiner anderen Prüfungsbehörde vorgelegt wurde.

Markus Schüler Berlin, 2011