Computational Analysis of Next-Generation Sequencing Data in Cardiac Function and Disease

Marcel Grunert

Juni 2012

DISSERTATION

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

eingereicht im Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutacher: Prof. Dr. Martin Vingron | Prof. Dr. Silke R. Sperling

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

Disputation: 22. August 2012

For my family

Preface

The first part of the thesis (Chapter 4) was published in the journal *PLOS Genetics* in 2011 entitled "The Cardiac Transcription Network Modulated by Gata4, Mef2a, Nkx2.5, Srf, Histone Modifications, and MicroRNAs" [1]. The small RNA read mapping tool MicroRazerS and its evaluation in Chapter 3 and 4 appeared in the journal *Bioinformatics* in 2009 with the title "MicroRazerS: Rapid alignment of small RNA reads " [2]. The last part about genomic sequence alterations, gene expression and microRNA profiling in patients with Tetralogy of Fallot has not been published yet, but a manuscript describing the oligogenic basis of isolated Tetralogy of Fallot is under review[3]. A follow-up manuscript about the impact of the found sequence variation on gene expression as well as an integrative analysis of gene and microRNA expression profiles in patients with Tetralogy of Fallot is in preparation.

The full study, which is described in parts in Chapter 4, integrates mRNA profiles with DNA-binding events of key cardiac transcription factors (Gata4, Mef2a, Nkx2.5, Srf), activating histone modifications (H3ac, H4ac, H3K4me2 and H3K4me3) and microRNA profiles in wildtype and siRNA-mediated knockdown. My contribution to this paper was the analysis of the ChIP-seq (Srf and H3ac) and microRNA-seq (after Srf knockdown) data. This includes the ChIP-seq read mapping and peak calling as well as the small RNA read mapping. I was involved in the analysis of overlapping transcription factor binding sites between ChIP-chip and ChIP-seq. I contributed in the analysis regarding the influence of H3ac and Srf marks on gene expression in the Srf knockdown. I compared microRNA expression profiles in HL-1 cells (cardiomyocytes cell line) to human normal hearts described in Chapter 2.2.1. I also constructed the Srf centered transcription network. For MicroRazerS, I was involved in the development and I did the evaluation as well as successful application of MicroRazerS to human and mouse small RNA-seq data. For the study described in Chapter 5, I carried out the complete computational analysis of targeted resequencing, mRNA-seq and microRNA-seq data.

I was further involved in the statistical assessment.

Acknowledgements The research presented in this thesis was carried out in the Cardiovascular Genetics group at the Max Planck Institute for Molecular Genetics. I want to thank all people who have helped and supported me and made my PhD studies a pleasant time.

First and foremost, I would like to thank my advisor Silke R. Sperling for the opportunity to pursue this research. I am grateful for her ideas, support and fruitful discussion throughout the years. She provided a creative and open environment to work in, with opportunities to attend international conferences to gain more insight into the exciting field of bioinformatics.

I also would like to thank Martin Vingron for supervising my PhD thesis and for supporting the issues of the PhD students in the institute.

All the present and former members of the Sperling group: I would like to thank Markus Schüler for several scientific discussions and about different things in everyday life, the time together on many conferences and his social being and kindness. It was a pleasure to work with him. I also like to thank Cornelia Dorn for her never-ending patience with ever-changing "lists", quite revealing discussions about biological questions and her wonderful way to deal with people, not least with a smile. I want to thank Jenny Schlesinger for her dedication in the lab, the accuracy in the evaluation of lab results and last but not least our shared rides back home, often rich on awesome conversations. I further thank Ilona Dunkel for all the work in the lab (in particular the amazing library preparation) and her unique and powerful being, which I admire deeply. I also thank Martje Tönjes for the initial motivation and her given (unforgettable) yoga sessions. I like to thank Barbara Gibas for the office and travel support and the never-ending supply with candies. I thank Siegrun Mebus for collecting patient tissue samples. I also would like to thank all other members of Sperling group for the unique research environment, namely Qin Zhang, Huanhuan Cui, Katherina Bellmann, Kerstin Schulz and Vikas Bansal.

I like to thank Silke Stahlberg and Yves Clement for the co-work in the PhD student association of the institute. I further would like to thank Anne-Katrin Emde for the collaboration in the development of MicroRazerS and the interesting discussions about NGS-based analysis approaches. I also want to thank Hugues Richard and Mar-

cel Schulz for their help to make the POEM algorithm more applicable for large-scale datasets and their discussions about statistical methods for RNA-seq data.

I am deeply grateful to all TOF patients and family members participating in the studies, and to all our collaborators and co-authors in the papers. In particular, I am also indebted to Markus, Jenny and Conny for great comments from proofreading this thesis.

Finally, my special thanks go to my mother, my brother and my girlfriend Daggi for their encouragement, love and support through all the years. All my achievements would not have been possible without you. Thank you so much. I love you.

Contents

List of Figures

Chapter 1

Introduction

1.1 DNA, Gene Expression and MicroRNAs

The genetic information in the deoxyribonucleic acid (DNA) is stored as a sequence of bases (or nucleotides) and the order of the nucleotides determines the genetic information. Each DNA strand consists of the four nucleotides adenine (A), cytosine (C), thymine (T) and guanine (G), arranged in a double helix. In eukaryotes, DNA is organized into chromosomes and located in the nucleus of each cell. For example, in human there are 23 chromosome pairs (one of each pair from both the mother and the father) representing approximately 3 billion base pairs, giving a total of 46 chromosomes per cell. The central dogma of molecular biology states the flow of genetic information in biological systems: Coding regions of DNA are transcribed into ribonucleic acid (RNA), which is then translated into proteins[4]. Unlike DNA, most RNA molecules are single-stranded and the nucleotide thymine is replaced by uracil (U), which differs from thymine by lacking a methyl group.

A gene, a contiguous region of DNA, corresponds to one transcribed unit which in turn is translated to one or more chains of amino acids called polypeptides of related or different functions[5]. When a gene is activated, the DNA strands separate and one of them serves as a template for copying a messenger RNA (mRNA). The complete gene region is first transcribed into a precursor mRNA (pre-mRNA) molecule that consists of coding exons alternating with non-coding introns, which is subsequently spliced into mRNA (whose sequence encodes the polypeptide).

In eukaryotes, transcription of protein-coding genes is carried out by RNA polymerase II (Pol II), a complex of 12 different proteins. A number of proteins is crucial for success-

ful localization of Pol II to the transcription start sites (TSS) and mRNA transcription including the general transcription factor and co-factors[6,7]. General transcription factors like TFIIB or TFIID (complexes consisting of the TATA-binding protein and other associated factors) bind in close proximity to the TSS and are involved in the separation of the DNA strands as well as the recruitment of Pol II. In addition to the general transcription factors, sequence-specific DNA binding transcription factors (TFs) can regulate gene transcription by interacting with this core transcriptional machinery[7,8]. They bind to one or multiple factor specific cis-regulatory elements located on the DNA called transcription factor binding sites (TFBS). These TFBS can be found in the core and proximal promotor regions (directly up- and downstream to the TSS), within exons and introns, in 5' and 3' untranslated regions of mRNAs, and even as far as 10 kilobases (kb) away from the respective genes[9,10]. Based on the regulatory function of TFs, cis-acting elements are classified into enhancers (activator) or silencers (repressor)[11].

The ability of transcription factors to bind cis-regulatory elements is highly dependent on their accessibility. In all eukaryotic cell nuclei, DNA is highly condensed into a structure called chromatin by the use of highly conserved proteins known as histones. Histones form complexes comprising two of each of the four core histone proteins (H2A, H2B, H3, and H4). The DNA double-helix is wrapped in approximately 1.75 turns around such an histone octamer to form the nucleosomes, which are connected through short linker-DNA of different length that is stabilized by the so-called linker histone protein H1[12]. Highly condensed DNA regions (heterochromatin) aggravates or hinders the binding of TFs and therefore they are associated with inactive gene transcription. On the other hand, less condensed or open DNA regions (euchromatin) are easily accessible by the transcriptional machinery and thus associated with an active transcription of genes. Changes and thus the associated degree of condensation in chromatin structure are dynamically controlled by epigenetic mechanisms like chromatin remodeling[13], DNA methylation[14] and histone tail modifications[15].

Covalent post-translational modifications of histone tails including acetylation, methylation, phosphorylation, ribosylation, sumoylation and ubiquitination can e.g. influence the wrapping of DNA around the histone octamer and thereby lead to an altered transcriptional accessibility[16]. For example, acetylation is a modification that neutralizes positively charged histone tails and therefore lowers the electrochemical coupling between the histone octamer and the wrapped DNA, which is thus more accessible[17]. This modification is catalyzed by a group of enzymes called histone acetyltransferases

(HATs). Histone acetylation, which is associated with an increased transcriptions, can be reversed by histone deacetylases (HDACs), which in turn are associated to a decreased expression level. Consequently, the interplay between HATs and HDACs activities regulates histone acetylation levels in the cells[18,19].

After transcription, gene expression is further controlled on a post-transcriptional level by RNA binding proteins. These proteins regulate RNA splicing, RNA processing, nuclear export and nuclear degradation. RNA splicing is the process that removes intron sequences from the pre-mRNA. Introns usually contain a clear signal for splicing, namely short sequences called splice sites[20]. Alternative splicing is the mechanism by which a pre-mRNA molecule produces different mRNA variants, by skipping, including, extending or shortening exon sequences, or retaining intron sequences. Besides alternative splicing, polyadenylation of pre-mRNA molecules and differential promoter usage can produce multiple transcript isoforms whose respective expression levels are regulated in a spatial and temporal manner. It has been estimated that 75-92% of human genes give rise to multiple isoforms[21,22]. The cellular abundance of mRNA molecules is of particular importance as it regulates the rate of protein synthesis. Besides transcription, RNA degradation directly determines the amount of mRNA. One way to regulate the decay of a mRNA molecule is the shortening of its poly-A tail, consisting of a series of adenine nucleotides, by specialized exonucleases.

MicroRNAs (or miRNAs) are short, single-stranded RNA molecules, usually ranging from 19 to 25 nucleotides (nt) in length, which regulate expression of target genes and thereby play an essential role in many biological processes. The first miRNA lin-4 was discovered in 1993 in *Caenorhabditis elegans* (C.elegans), and seven years later a second one, let-7, was found to regulate later developmental stages of C. elegans in a similar manner to lin-4[23,24]. It was soon realized that both lin-4 and let-7 were evolutionarily conserved in the genomes of eukaryotes, implicating a more universal role for these small RNA molecules[24,25]. Since then, several hundreds of miRNAs present in both plant and animal genomes were revealed[25–32]. Yet until now only a limited number of these have been characterized in depth[33]. Most miRNAs are found in intergenic regions and contain their own miRNA gene promoter and regulatory units[25,27,28]. Approximately 40% of the miRNAs lie in introns of protein and non-protein coding genes, or even in exons[34]. These miRNAs are usually found in sense orientation and thus they are regulated together with their host genes[34–37].

Generally, a miRNA is transcribed into a RNA hairpin loop by RNA Polymerase II

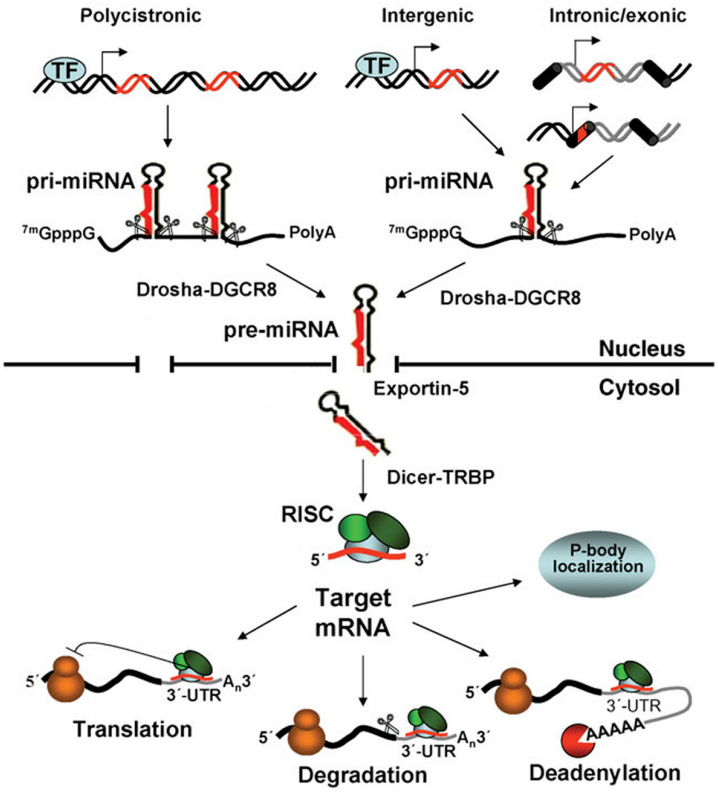

Figure 1.1: Genomic organization, biogenesis and function of miRNAs. Figure taken from Fazi and Nervi[38] and modified.

or III and capped to form the primary miRNA transcript (pri-miRNA). The following cleavage of the pri-miRNA by ribonucleases Drosha and DGCR8 in the nucleus yields a stem-loop structure of approximately 70-100 nucleotides. This precursor hairpin (pre-miRNA) is transported in the cytoplasm by Exportin-5, where it is further cleaved by the Dicer protein into the miRNA/miRNA* duplex (Figure 1.1). The guide strand of the miRNA is then loaded together with Argonaute (Ago2) proteins into the RNA-induced silencing complex (RISC). After assembly of the RISC complex, the mature miRNA binds to a short recognition sequence at the 5' end, the so-called seed region, to mRNAs with complementary sequence in their 3' untranslated region (UTR) usually resulting in mRNA cleavage, translational repression or mRNA degradation. The passenger strand, the minor product denoted by a star, is commonly degraded, though this is not always the case[39]. Both strands of the miRNA/miRNA* duplex can potentially act as a functional miRNA, but only one is finally incorporated into the RISC complex.

As mentioned before, miRNAs usually silences their target mRNAs and 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 target mRNAs[40–42]. Moreover, miRNAs may themselves be mediators of default repression[40], also suggested by the growing evidence for a high abundance of miRNAs in the cell. The current release (v.18) of miRBase[43], the primary online repository for miRNA sequences and annotation, contains over 18,000 hairpin precursor miRNAs, expressing over 21,600 mature miRNAs, in 168 species. The database was established in 2002 with 218 entries and the number of miRNA sequences deposited in miRBase has risen approximately exponentially - in the last 3 years the number has almost tripled[43]. For human, the current miRBase version contains 1,527 hairpin precursor miRNAs and 1,921 mature miRNAs. Further, over 60% of all human protein-coding genes are predicted to be regulated by miRNAs[44], with one miRNA regulating hundreds of mRNAs each[45,46].

1.2 Next-Generation Sequencing

The primary method of sequencing DNA referred to as chain-termination sequencing is commonly known as Sanger sequencing[47]. It was first developed by Frederick Sanger in 1977[48]. Driven by the goal of deciphering complete gene sequences (later entire genomes like the human) and based on the associated throughput requirements of DNA sequencing, the Sanger method has almost exclusively been carried out with semi-automated capillary electrophoresis (Figure 1.2A)[49,50]. Moreover, the semi-automated implementations of the Sanger biochemistry has become the 'gold standard' in terms of both sequence read length (up to ∼1,000 bp) and sequencing accuracy (per-base 'raw' accuracy as high as 99.999%)[50,51]. In the last years, various second generation or, more commonly, next-generation sequencing (NGS) technologies have been developed, which will be still rapidly further developed. These parallel processing techniques are able to generate several orders of magnitude more sequence output and have significantly reduced the cost of DNA sequencing compared to conventional Sanger sequencing. Although the technologies differ in their biochemistry they all follow the principle of cyclic-array sequencing, where a dense array of DNA features is sequenced by iterative cycles of enzymatic reactions combined with imaging-based data detection (Figure 1.2B).

The different NGS technologies have been released as commercial products, with the most popular being the Solexa Genome Analyzer (Illumina), the 454 Genome Se-

Solexa	Bridge amplification Polymerase-based sequencing-by-synthesis Read length: 36 to 150 bp Error rate: 1% per bp Error rate increases preferentially at the 3' end of reads Dominant error type: substitutions	50–54
454	Emulsion PCR Polymerase-based pyrosequencing Relatively long read length: 250 to ≥400 bp Error rate: 0.5% per 250 bp and 0.1% per 400 bp Dominant type of error: insertions or deletions Long single dNTP strings (homopolymer repeats) unreliable (8 bp linearity)	50–52,55–57
SOLiD	Emulsion PCR Ligase-based sequencing (octamers with two-base encoding) Read length: 50 to 75 bp Low error rate: <0.1% per bp Dominant error type: substitutions (colour shift) Two-base encoding provides inherent error correction	50–52,58
Helicos	Single molecule Polymerase-based sequencing (asynchronous extension) Read length: 25 to 55 bp (35 bp in average) Error rate: <1% per bp (Substitution 0.2%, Insertion 1.5%, Deletion 3.0%) Dominant error type: deletions No PCR amplification (high reproducibility)	50–52,59,60

Table 1.1: Next-generation sequencing technologies with they current properties. For all platforms single- and paired-end read sequencing modes are available. The indicated lengths are for single reads.

quencers (Roche Applied Science), the SOLiD platform (Applied Biosystems) and the HeliScope Single Molecule Sequencer technology (Helicos). There are important differences among these platforms themselves that result in advantages but also in disadvantages with respect to specific applications (Table 1.1). Some applications, e.g. genomic resequencing, are more tolerant regarding short sequence fragment (read) lengths than others such as *de novo* assembly. For applications relying on counting sequence tags, e.g. the quantification of protein-DNA interactions, the given amount of sequencing should be split into as many reads as possible, whose length are above some minimum that allows the exact placement to a reference. In general, a high number of reads provides greater depth and therefore, sequence confidence. Finally, the overall accuracy as well as specific error distribution of individual technologies, such as the propensity for systematic errors, are also important[50].

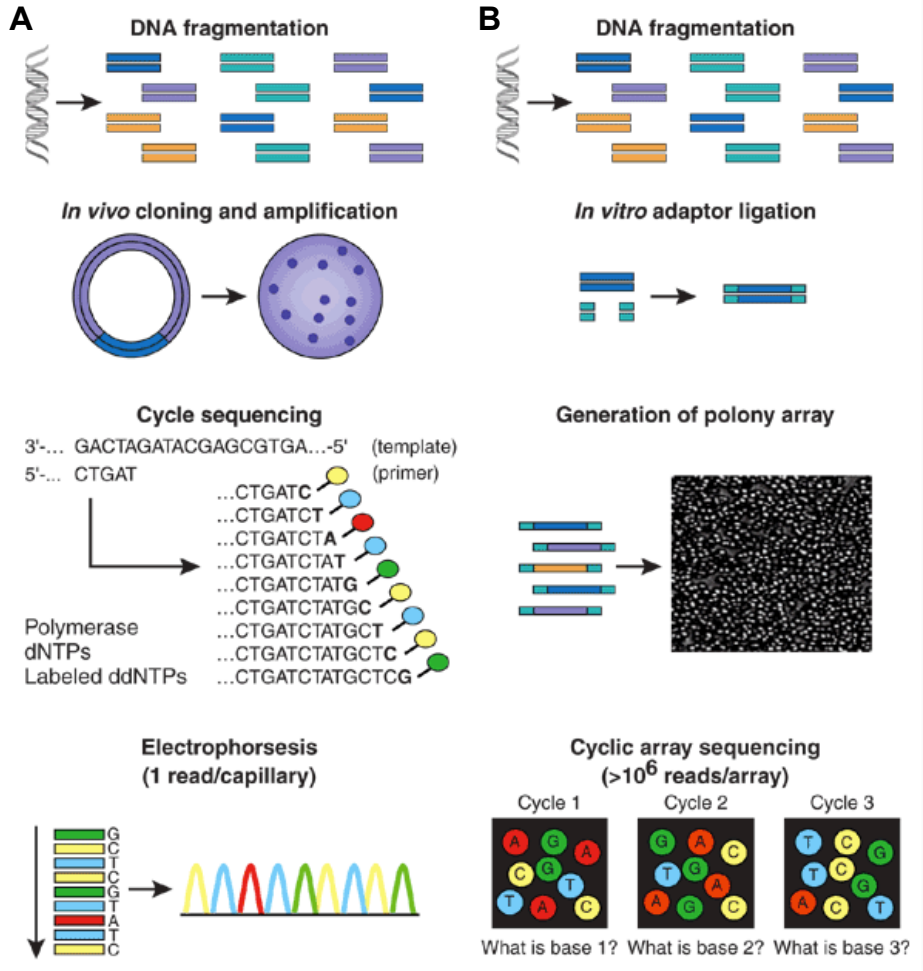

Figure 1.2: Sanger sequencing compared to next-generation sequencing. (A) With high-throughput shotgun Sanger sequencing, genomic DNA is fragmented and afterwards cloned into a plasmid vector and transformed into bacteria (e.g. E. coli). A single bacterial colony is selected and the plasmid DNA is isolated. Each cycle sequencing reaction generates a ladder of dye-labeled products, which are subjected to high-resolution electrophoretic separation in one sequencing run. The fluorescence labeled fragments of discrete sizes pass a detector generating a four-channel emission spectrum, which is finally used for the sequencing trace. (B) In next-generation shotgun sequencing methods, common adaptors are ligated to fragmented genomic DNA, which is then treated to create an array of millions of immobilized PCR colonies, called polonies. Each polony contains many copies of a single shotgun library fragment. In cyclic reactions, sequencing and imaging-based detection of fluorescence labels build up a contiguous sequencing read for each polony. Figure taken from Shendure and Ji[50].

One key finding of the Human Genome Project is that any two human individuals

are nearly 99.9% identical in their genomic DNA sequence. The residual 0.1% leads to several million differences, with some of these variations giving rise to certain diseases, drug responses and other complex phenotypes. The first differences observed in the human genome were mainly rare changes in the quantity and structure of chromosomes, called structural variants. Structural variants are genomic alterations that involve segments of DNA that are usually larger than 1 kb, and can be microscopicly detectabled. There are different types of structural variations including copy number variations (CNVs), segmental duplication or low-copy repeat (LCR) and chromosomal rearrangements such as inversions and translocations. CNVs are alterations that result in a copy number change of one or more sections of the DNA including duplications, insertions and deletions. A CNV that occurs in more than 1% of the population is called a polymorphism.

In addition to structural variations, there are smaller and more abundant alterations. Such local variations include single nucleotide variations or polymorphisms (SNVs or SNPs, respectively) as well as small (usually <50 bp) InDels, which includes both insertions, deletions, and the combination thereof. A SNP, a variation at a single site in DNA, is the most frequent type of genetic variations in the (human) genome. They are highly conserved within a population and make excellent genetic markers. Early estimates predicted that there are at least 10 million SNPs within the human population[61], meaning that SNPs occur in 1 of 300 base pairs, on average, among the ∼3 billion base pairs of the human genome[62]. Due to the efforts of the 1000 Genomes Project the number of known human SNPs currently exceeds 35 million[63]. In addition, there are around 100-200 novel mutations (single base changes) in the human genome per generation. This is equivalent to one mutation in every 30 million base pairs. Most of these are benign and have no apparent effect on the health or phenotype, and only very few mutations are accumulated over several generations, which can lead to certain diseases[64].

Alleles are forms of a gene, which are located in the DNA of an organism. The human genome has thousands of genes with different sets of alleles and not necessary all genes will only have two possible alleles since there are only two homologous chromosomes for a diploid organism. For example, in human blood types there are three possible alleles (i.e. A, B and 0). Alleles are often composed of one or more SNPs and therefore, the most commonly called base, which is not the reference base, for a given position in the reference based sequence alignment is often defined as alternate allele. If the alter-

nate allele frequency is between 20% and 80%, the genomic position is usually called as a heterozygous variation, and homozygous if the frequency is over 80%[51]. Further, the minor allele frequency (MAF) is the ratio of chromosomes in the given population carrying the less common (rare) variant to those with the more common variant. By definition the MAF is less or equal to one.

1.3 The Human Heart and Congenital Heart Disease

The heart is one of the most important organs in the human body. It pumps the blood, which is essential for nutrition and oxygen supply of all living cells, throughout the body by repeated, rhythmic contractions. In human, deoxygenated blood from the body is transported through the venae cavae (superior and inferior, respectively) into the right atrium, though the tricuspid valve into the right ventricle, and finally through the pulmonary valve into the pulmonary artery and further into the lung. The oxygenated blood returns from the lung into 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 back into the body (Figure 1.3). The heart is the first organ to form and function during embrogenesis and starts beating after 20 days of gestation in human[65]. The complex development of the heart involves the spatial and temporal orchestration of various molecular pathways and complex morphogenetic changes, which are precisely controlled by an evolutionarily conserved gene program. The mammalian cardiogenesis requires a diverse set of cell types including cardiomyocytes, cells of the conduction system, smooth muscle cells, endothelial and valvular cells[66]. The formation of these various cardiovascular cell lineages has its basis in the existence of a closely related set of multi-potent progenitors in the early embryonic heart field, which can be divided into the primary heart field (or first heart field; FHF) and secondary heart field (SHF)[67,68]. After around eight weeks of gestation the four-chambered human heart is completely developed[69], the left ventricle 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[70].

The molecular network underlying cardiogenesis is evolutionarily conserved from simple model organisms to higher vertebrates and comprises regulatory interactions between numerous transcription factors, their downstream target genes and upstream signaling pathways[72]. A core set of conserved DNA-binding transcription factors, including Gata, Hand, Nkx2, Mef2, Tbx-factors and Srf, regulates heart development in a decisive

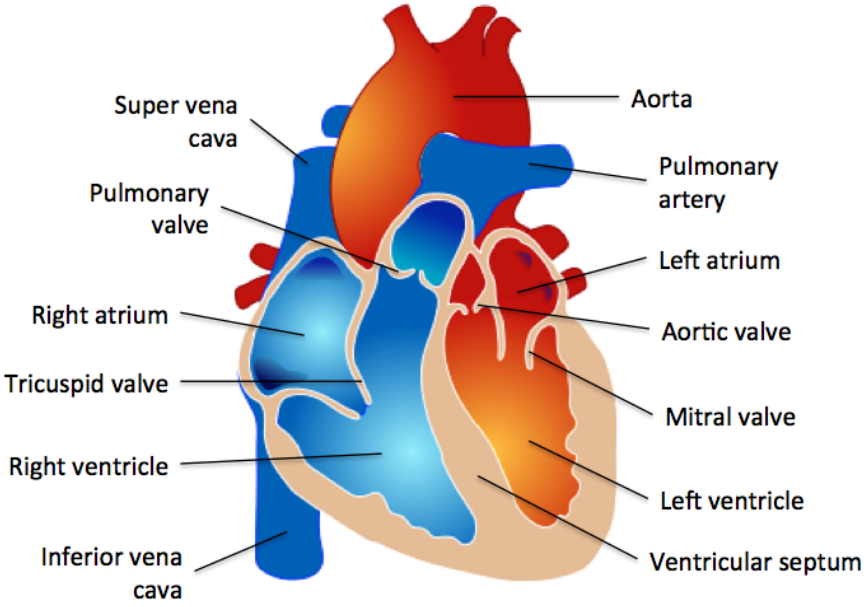

Figure 1.3: Schematic representation of the mature four-chambered human heart. Oxygen-rich blood is indicated in red and oxygen-poor blood is indicated in blue. Figure taken from M. Ruiz[71] and modified.

manner. They play pivotal roles for the differentiation, maturation and homeostasis of cardiomyocytes and can directly interact providing cooperative regulation of individual target genes. For example, the homeobox transcription factor Nkx2.5 physically interacts with Gata4 and Tbx5 to synergistically activate several downstream target genes[73]. The zinc-finger transcription factor Gata4 on the other hand can physically interact with Hand2[74], Nkx2.5[75], Mef2[76], Tbx5[77] and Srf[78].

The widely expressed serum response factor (Srf) is important for heart and muscle development and is well-known to bind to the CArG-box motif [CC(A/T)$_6$GG], a DNA consensus sequence, in promotors of its target genes[79] and moreover auto-regulates its own expression[80]. Srf is involved in the regulation of the cell cycle, apoptosis, muscle cell differentiation and cellular growth, as well as in the actin cytosketon. It regulates the expression of structural muscle genes such as actins and myosins, which belong to the contractile apparatus[78,79,81–83]. Furthermore, Srf is known to interact with both positive and negative co-regulators. For example, together with Gata4 and Nkx2.5, Srf directs early cardiac gene activity[84]. Since Srf is ubiquitously expressed, it alone cannot account for smooth muscle-specific gene expression but through the association with

e.g. Myocardin (Myocd), a smooth muscle and cardiac muscle-specific transcriptional co-activator, it can activate muscle gene expression[82,85].

All these transcription factors also regulate each others expression, thereby reinforcing and stabilizing the cardiac gene program[72,86]. For example, the cardiac T-box factor Tbx20 interacts with Gata4 to activate both Mef2c and Nkx2-5 enhancers[87]. However, they do not regulate on direct transcriptional level, but indirectly by influencing the chromatin status of their target genes. For example, Mef2 proteins can act as transcriptional activators and repressors through the interaction with HATs and HDACs, respectively[88,89]. It has been reported that the Srf-cofactor Myocardin recruits the HAT p300 to Srf binding sites whereby histone 3 acetylation (H3ac) is induced and gene expression enhanced[90]. The HAT p300 not only acetylates lysine residues on histone 3 but also on Gata4, thereby enhancing its DNA-binding and its activating potential[91]. Further, Srf as well as Gata4, Mef2c and Nkx2.5 are negatively regulated by interaction with HDAC4, a transcriptional repressor of muscle gene expression[92].

Like the interaction between genetic and epigenetic factors, miRNAs are interacting with all regulatory levels, leading to complex regulatory networks that maintain correct cardiac morphogenesis. For example, Srf regulates the transcription of miRNAs such as the smooth muscle relevant miR-143 and miR-145[93]. Feedback loops between Srf/Mef2 and muscle-specific miR-133/miR-1 have been described and both miRNAs are expressed throughout heart development playing important roles in muscle proliferation and differentiation[94–97]. Furthermore, miR-1 promotes myogenesis by targeting HDAC4[95] and thus represents a connection to histone acetylation. The loss of function of any of these transcription factors, their cofactors or miRNAs can dramatically affect the regulatory cascades with consequences for cardiovascular development and congenital heart disease.

Congenital heart disease (CHD) are the most common birth defects in human with an estimated incidence of around 1% in all live births[98]. They range from minor or even subclinical defects to complex malformations. Due to the significant advances in cardiac care with regard to cardiac surgery and interventions, the mortality of congenital heart disease has significantly reduced over the last decades. Recently it was estimated that nearly 760,000 individuals with CHD born after 1990 will be alive by the year 2020[99]. Almost all parts of the heart can be affected and the disease phenotype can be classified into septation, left-sided obstruction and cyanotic heart defects[70]. Septation defects are e.g. the atrial septal defect (ASD), the ventricular septal defect (VSD) or

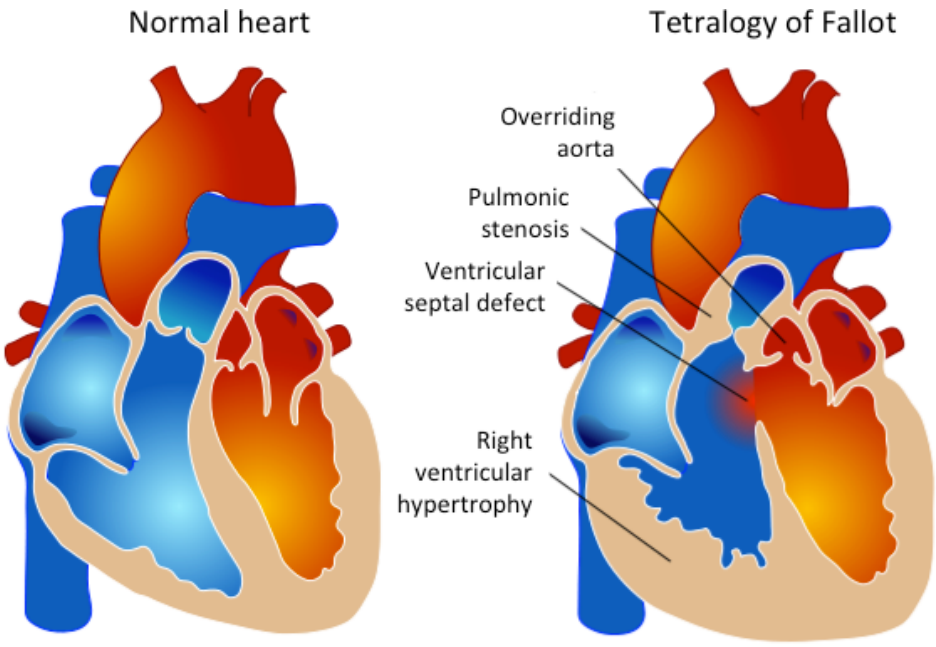

Figure 1.4: Schematic representation of a normal human heart (left) and a heart with the 'Tetralogy of Fallot' phenotype (right) depicting the four clinical features. Oxygen-rich blood is indicated in red and oxygen-poor blood is indicated in blue. Figure taken from M. Ruiz[71].

the atrioventricular septal defect (AVSD). Typical, left-side obstruction defects are the aortic stenosis and an interrupted aortic arch. Cyanotic heart defects result from the mixing of oxygenated and deoxygenated blood and cause a blue skin color, also referred as "blue baby syndrome". Examples for such defects are a transposition of the great arteries (TGA), tricuspid atresia, Ebstein's anomaly of the tricuspid value, the persistent ductus arteriosus (PDA) and Tetralogy of Fallot (TOF). TOF is the most common form of cyanotic defects and if untreated it ultimately leads to cardiac failure with a survival rate of around 60% after four years[100]. It is a complex disease with four distinct clinical features: A VSD, a right ventricular outflow track obstruction (narrowing at or just below the pulmonary valve), a right ventricular hypertrophy (thickening of the right ventricular wall) and an overriding aorta, a biventricular origin of the aortic valve (Figure 1.4).

Approximately 20-30% of CHD occur in association with other birth defects as part of a syndrome, such as DiGeorge syndrome or Holt-Oram syndrome, and in many of them chromosomal (e.g. loss of one copy of TBX1 in DiGeorge syndrome[101,102]) as well as gene mutations (e.g. TBX5 mutation in Holt-Oram syndrome[103,104]) could be

identified as causative for the defect. Only a minority of CHD are monogenic disorders that follow a clear Mendelian inheritance. Linkage analysis in non-syndromic families with Mendelian inheritance pattern identified several gene mutations in the etiology of human CHD such as ACTC1 (ASD[105]), GATA4 (ASD[106]), JAG1 (TOF[107]), MYH6 (ASD[108]), MYH11 (PDA[109]), NKX2.5 (ASD[110]), NOTCH1 (bicuspid aortic valve and aortic stenosis[111]) and ZIC3 (TGA[106]). However, the majority of CHD do not segregate in Mendelian ratios, although they show familial aggregation suggesting that genetic factors play a role in their development[112]. Some disease-associated mutations have been found in genes which control cardiac development including CITED2[112], GATA4[113], NKX2.5[114], NOTCH1[115], TBX1[116] and TBX20[117]. Nevertheless, the genetic mechanisms underlying non-chromosomal or non-Mendelian "sporadic" defects are poorly understood[118]. Typically, the gene mutations of sporadic CHD are individually unique, resulting in allelic heterogeneity[118]. Furthermore, mutations are always heterozygous and as in the case reported, the defects were transmitted by an unaffected parent, indicating that these rare mutations are incompletely penetrant[119].

Beside the genetic influence, it is long known that prenatal environmental factors such as alcohol, anti-depressants, anti-epileptic drugs, deficiency of zinc or vitamin A, herbicides, diabetes, obesity or infection like rubella significantly enhance the probability of CHD[120–126]. One of the first publications regarding the etiology of CHD was introduced by James Nora in 1968. Nora already proposed CHD to be multifactorial disorder caused by genetic and environmental influences[127]. In 1976, he published the first study showing a familial recurrence risk of 2-5%, which underlines the genetic background but clearly points to additional factors[128]. Our current understanding is that of an oligo- or multigenic background (i.e. 3-8 or even more mutations). There are international projects ongoing (e.g. HeartRepair, CardioGeNet or CHeartED), which study genomic variations at a large scale in several cohorts of CHD. With respect to preliminary data, CHD are most likely caused by a panel of genetic variations. At least a subset of these mutations is inherited from parents. Probably each mutation only modestly effect protein function or expression and manifestation of the disease occurs only when combined with additional genetic, epigenetic (miRNAs, histone modifications or DNA methylation changes) or environmental insults. Epigenetic mechanisms might represent the mechanism by which environmental factors impact on the disease and its trans-generational transmission.

1.4 Purpose and Aims

In the last years, next-generation sequencing has revolutionized almost all fields of genetics and has become the method of choice for genome analysis. The emergence of NGS platforms requires increasing demands on statistical methods and bioinformatic approaches for the analysis and the management of the huge amounts of sequence data generated in a very short time scale by these technologies. Moreover, there is a wide range of NGS applications, rapidly developing, making the computational analysis of their associated datasets very challenging.

This thesis aimed to develop novel computational approaches and bioinformatics tools for the analysis of NGS datasets generated within the group as well as publicly available and eventually answer biological questions regarding cardiac function and disease.

In human, a large number of transcription factors, different histone modifications and post-transcriptional regulators like miRNAs modulate the mRNA profile corresponding to thousands of protein-coding genes. However, we lack data showing interactions between these levels of regulation since in the past insights were obtained by focusing on each level independently. The first study in this thesis aimed to elucidate the combinatorial regulation of cardiac DNA-binding transcriptions factors (ChIP-seq of Srf) influenced by histone modifications (histone 3 acetylation) and regulatory miRNAs (miRNA-seq) in cell culture. To gain insight into the transcriptional regulation of cardiac mRNA profiles, the different modulators need to be viewed in context to each other.

Tetralogy of Fallot accounts for 7-10% of all congenital heart disease, which are the most common birth defect in human. CHD are most likely caused by a panel of genetic variations with each effecting expression or protein function only modestly and manifest as disease only when combined with additional genetic, epigenetic or environmental alterations. In the past, the discovery of oligo- or multigenic disorders has been less amenable to conventional genetic techniques. The second project aimed to identify the genetic basis of TOF performing a multilevel study comprising targeted resequencing of heart- and muscle-relevant genes and miRNAs in patients with TOF, parents and controls as well as whole transcriptome (mRNA-seq) and miRNome (miRNA-seq) analysis in TOF cases and healthy unaffected individuals using the latest NGS techniques.

Chapter 2

Next-Generation Sequencing Applications and Datasets

2.1 Applications

In the last few years, the application of semi-automated Sanger sequencing for the genome analysis has been replaced by next-generation sequencing (NGS) methods. The ability to sequence millions of DNA fragments in less than one day is the major advance offered by NGS. For gene expression analysis the conventional microarrays are now being replaced by sequenced-based methods, which can identify and quantify rare transcripts without prior knowledge of a particular gene. In summary, the huge amount of low-cost reads makes NGS technologies useful for several application. There is an impressive range of NGS applications, rapidly developing. This includes the sequencing of expressed mRNAs and miRNAs, the identification of genome-wide protein-DNA interactions such as transcription factor binding sites or chromatin histone mark, and the detection of sequence alterations. The applications and their associated datasets, computational analyzed in this thesis, are described in following.

2.1.1 Genome-wide Mapping of Protein-DNA Interactions

A powerful technique for genome-wide identification of protein-DNA interactions such as transcription factor binding sites[130,131] or chromatin histone marks[132,133] is chromatin immunoprecipitation (ChIP) followed by either microarray detection (ChIP-chip) or, more recently, next-generation sequencing (ChIP-seq). In a ChIP experiment, proteins and protein complexes are cross-linked to DNA via formaldehyde. Afterwards,

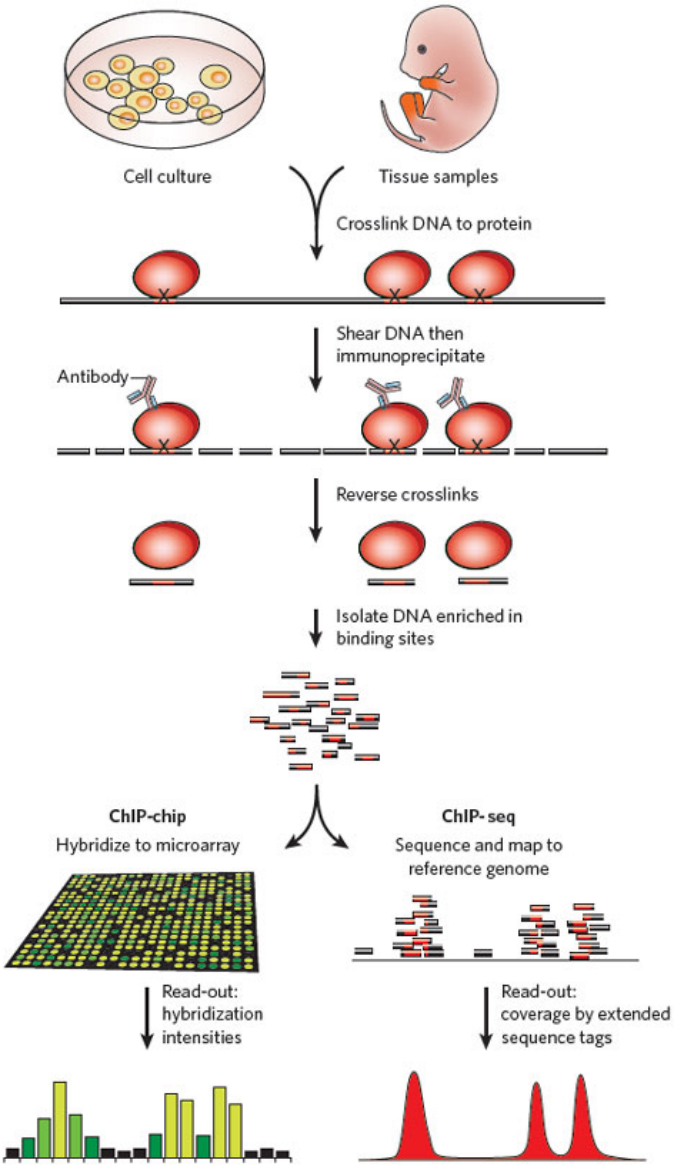

Figure 2.1: Schematic representation of a chromatin immunoprecipitation (ChIP) experiment followed by microarray detection (ChIP-chip) or next-generation sequencing (ChIP-seq). Figure taken from Visel *et al.*[129].

chromatin is shared by sonication (ultrasound) into small fragments, which are 200-600 bp in length[134]. In the next step, the DNA fragments bound to the protein of interest are enriched using an antibody specific to the protein. The DNA fragments which are not bound to the protein will be washed away. After reverse cross-linking

and purification of the DNA to remove the proteins, an enriched DNA sample called 'ChIP sample' is obtained. In many studies, an additional sample, also known as 'Input sample' or just 'Input', is prepared in parallel which is not immunoprecipitated to measure the experimental background. Finally, after size selection (typically in range of ∼150-300 bp[134]) and further processing (e.g. additional amplification if the amount of enriched DNA fragments is too low), the DNA fragments are determined to measure protein-DNA binding regions. Previously, ChIP-chip was the most common technique to study these protein-DNA interactions[135,136]. In ChIP-chip the enriched DNA fragments are hybridized to a microarray, e.g. genome tiling arrays for organisms with small genomes or custom designed arrays for certain regions of interest such as promoters for a selected number of genes. In the last few years, ChIP-seq, which combines ChIP with high-throughput massively parallel sequencing, is increasingly being used for mapping protein-DNA interactions *in vivo* on a genome-wide scale. In ChIP-seq, tens of millions of short DNA fragments, or sequence reads, are sequenced directly from both ends instead of being hybridized on an array. By computationally mapping these sequence reads to a reference genome and looking for genomic regions (peaks) where they are enriched, genome-wide mapping locations of protein-DNA interactions can be identified (Figure 2.1).

Compared to ChIP-chip, ChIP-seq offers several advantages. In general, it has higher resolution, fewer artefacts, greater coverage and a much broader dynamic range[134]. The main improvement is probably the base pair resolution. ChIP-seq provides single nucleotide resolution by measuring enrichment based on tag (read) counts whereas ChIP-chip measures enrichment by intensities of hybridization which may saturate at high signal, i.e. the intensity signal measured on arrays is not linear over its entire range. Moreover, the resolution in ChIP-chip is array-specific, generally in 30-100 bp range, and for example, high density tiling arrays require a large number of probes and are very expensive for large genomes[137]. In addition, ChIP-seq does not suffer from biases and noise caused by cross-hybridization including varying GC content, length, concentration or secondary structure of the target and probe sequence[134]. Further, only DNA fragments that are unique in the genome are spotted especially on the microarray which exclude highly repetitive regions which have already been shown to contain regulatory sites[138,139]. Moreover, only 48% of the human genome is non-repetitive, but using ChIP followed by next-generation sequencing 80% is mappable with 30 bp reads and 89% with 70 bp reads[140]. In addition, the fraction of reads that can be uniquely mapped to the genome decrease after ∼25-35 bp and is marginal beyond 70-100 bp[141].

Likewise ChIP-seq also has some disadvantages. For example, there are sequencing errors, especially towards the end of each read, although they have been reduced substantially as the technologies have improved. There is also a bias in GC-rich regions, both in library preparation and in amplification before and during sequencing [134,142,143]. Moreover, there is a loss of sensitivity and specificity in the detections of enriched regions when an insufficient number of sequence reads is generated [134]. Nevertheless, ChIP-seq has become the method of choice for almost all ChIP experiments, not only because of the rapidly decreasing costs of sequencing.

2.1.2 Quantification of Gene Expression and MicroRNA Profiling

The transcriptome is the pool of all transcribed elements in a given cell and RNA sequencing (RNA-seq) is a developed ultra high-throughput sequencing technology that enables researchers to discover, profile and quantify RNA transcripts across the entire transcriptome including mRNAs, non-coding RNAs and small RNAs [144–148]. RNA-seq provides in-depth information on the transcriptional landscape with unprecedented sensitivity and throughput [21]. It enables to outperform the previous sequence-based approaches starting with the analysis of expressed sequence tags (ESTs [149,149]) to high-throughput tag-based methods including serial analysis of gene expression (SAGE [150,151]), cap analysis of gene expression (CAGE [152]) and massively parallel signature sequencing (MPSS [153]).

In general, polyadenylated RNAs (poly(A)+) in a biological sample are extracted and converted into more stable cDNA fragments which are randomly sheared by either nebulization or sonication. After size selection, the fragments are amplified and adapters are ligated to one or both ends of the fragments. Finally, each fragment is sequenced using an NGS approach to obtain short reads from one end (single-end sequencing) or both ends (pair-end sequencing). Depending on the NGS technology, the reads are typically 30–400 bp in range [144]. There are several RNA-seq protocols varying in extracting mRNAs or other small RNAs like miRNAs (small RNA-seq or, according to miRNA profiling, miRNA-seq) as well as other non-coding RNAs, such as piwi-interacting RNAs (piRNAs) and short interfering RNAs (siRNAs). These small RNAs may be shorter than the sequenced reads and the sequencing process can reach into the adapter. As a consequence, the ends of the reads may contain variable lengths of adapter sequence. For example, miRNAs and siRNAs are ∼21-23 nucleotides in length and piRNAs are ∼25-35 nucleotides long whereas the minimum read lengths of the

different NGS technologies are usually longer (Table 1.1). In addition, small RNAs can be directly sequenced after adapter ligation, larger mRNAs must be fragmented into smaller fragments (∼200-500 bp) to be compatible with most of the NGS technologies[144]. Another key consideration in the library construction is whether or not to prepare strand-specific libraries[148]. The basic RNA-seq protocol is not strand-specific, meaning that the orientation of the reads is lost. The orientation is important for the annotation, especially for regions with overlapping genes from opposite directions.

2.1.3 Targeted Resequencing of Genomic DNA

Whole-genome sequencing of complex organisms such as human allows to gain a deeper understanding of the full range of genetic variations and to define the role of such sequencing routine in phenotypic variations as well as the pathogenesis of complex traits[154]. However, due to high costs and time exposure it is not yet feasible to sequence complex genomes in their entirety. For example, to obtain a 30-fold coverage of the full human genome, 90 Gb (gigabases) must be sequenced. Consequently, target enrichment methods have been developed, in which genomic regions of interest are isolated from a DNA sample before sequencing, focusing on these targets and their genomic variations. Targeted resequencing of genomic DNA is more time- and cost-effective. The resulting data are considerable less costly to analyze[154]. Furthermore, target sequencing has been shown to detect variants that are missed by whole-genome sequencing, suggesting that deep-targeted sequencing affords greater sensitivity than even genome coverage[155].

Several methods for target enrichment are available[157–160]. The approach used in this study relies on an array-based hybridization capture method[161–163]. This technology was first adapt to be compatible with next-generation sequencing by Roche NimbleGen. As a first step, a sequence capture array is made against that target regions in the genome. For example, NimbleGen sequence capture arrays are available that capture up to 5 Mb (385K array) or up to 50 Mb (2.1M array). Afterwards, a shot-gun sequencing library is build from genomic DNA by sonication or nebulization and hybridized to the sequence capture array. The unbound fragments are removed by washing and the enriched fragments are eluted and recovered from the array. The enriched fragments are then amplified by ligation-mediated polymerase chain reaction (LM-PCR) and the success is measured by quantitative PCR (qPCR) at control loci. Finally, a sequencing library enriched for target regions is ready for high-throughput sequencing

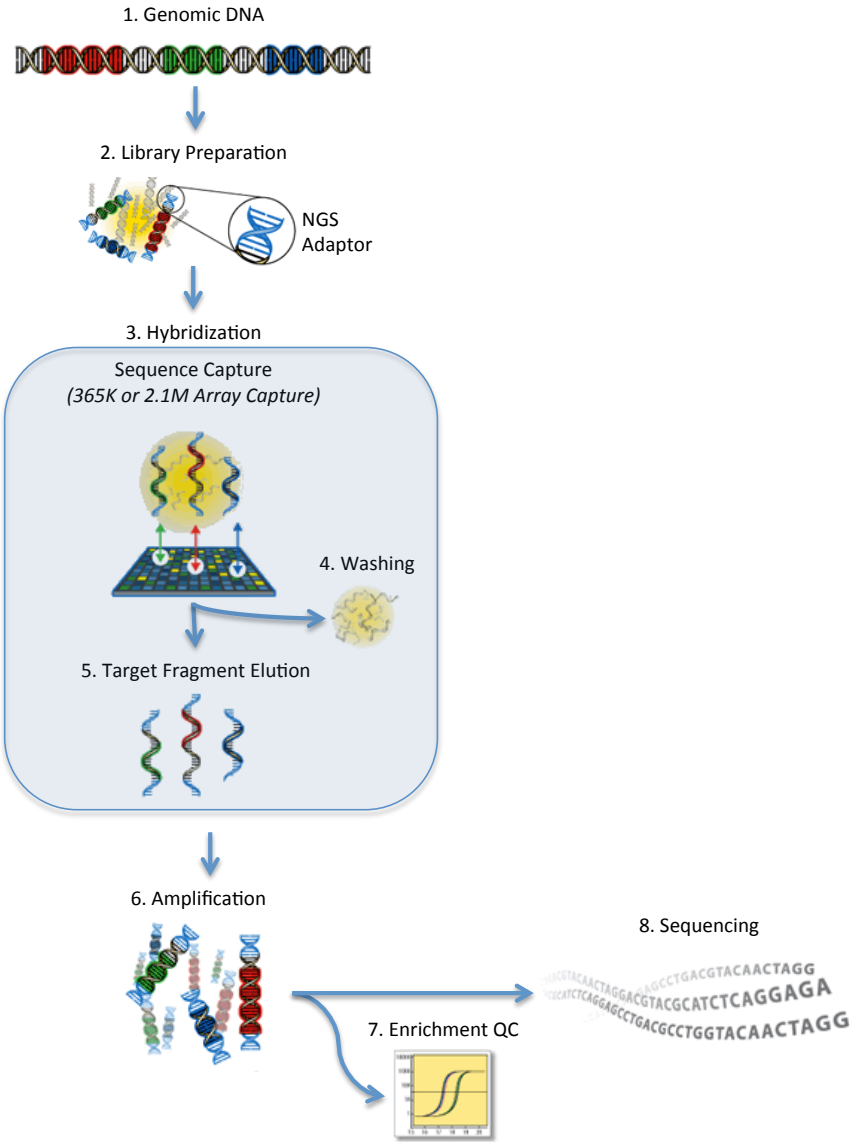

Figure 2.2: NimbleGen Sequence Capture technology for the enrichment of genomic target regions from genomic DNA. Figure taken from Roche NimbleGen[156] and modified.

using the Roche 454 Genome Sequencer (Figure 2.2)[156]. Just recently, modifications and optimizations of the original protocol enables the usage of the Illumina Genome Analyzer[154].

2.2 Datasets

The experimental datasets described in the following were generated in the group of Silke R. Sperling (*Cardiovascular Genetics*) at Max Planck Institute for Molecular Genetics. All NGS datasets have been computational analyzed in this study (see Chapter 4 and Chapter 5 for results). The experiments were conducted to study individual components of the transcriptional regulatory network of the vertebrate heart. Moreover, considering the opportunities of next-generation sequencing technologies, we aimed to gain deeper insights into the genetic causes of congenital heart disease. Next-generation sequencing was performed by the group of Bernd Timmermann (*Next Generation Sequencing Service*) at the Max Planck Institute for Molecular Genetics and by ATLAS Biolabs GmbH.

2.2.1 ChIP-seq Data of Srf and Histone 3 Acetylation in Cell Culture

The murine cardiomyocyte cell line HL-1 was used in all ChIP experiments described in the following. This cell line is a feasible model to study cardiomyocytes, as mRNA and miRNA expression profiles obtained from HL-1 cells are highly comparable to the one observed in mouse hearts right after birth (Pearson correlation coefficient of 0.95, Figure 2.3A) and human right ventricle (Pearson correlation coefficient of 0.90, Figure 2.3B). See Schlesinger *et al.*[1] for more information regarding the data and its comparison.

To study cardiac regulatory networks the initial step was the observation of the binding of the key transcription factors Gata4, Mef2a, Nkx2.5 and Srf to promoters of target genes using ChIP-chip. These transcription factors play pivotal roles for the differentiation, maturation and homeostasis of cardiomyocytes. The ChIP experiments were performed and previously analyzed in our group (see Schlesinger *et al.*[1]). With the focus on Srf, several hundreds of transcription factor binding sites could be identified (in total 1,335), which were related to 1,150 Srf target genes[1]. In addition, ChIP-chip data regarding the four activating histone modifications histone 3 acetylation (H3ac), histone 4 acetylation (H4ac) and histone 3 di- and trimethylation (H3K4me2/3) was used. These four histone modifications were described to promote an open chromatin state[164–167] and were generated and previously analyzed in our group also using ChIP-chip techniques and linear modeling[168]. With the focus on H3ac, 3,453 target genes were defined to be associated to 3,210 H3ac peaks in ChIP-chip[168].

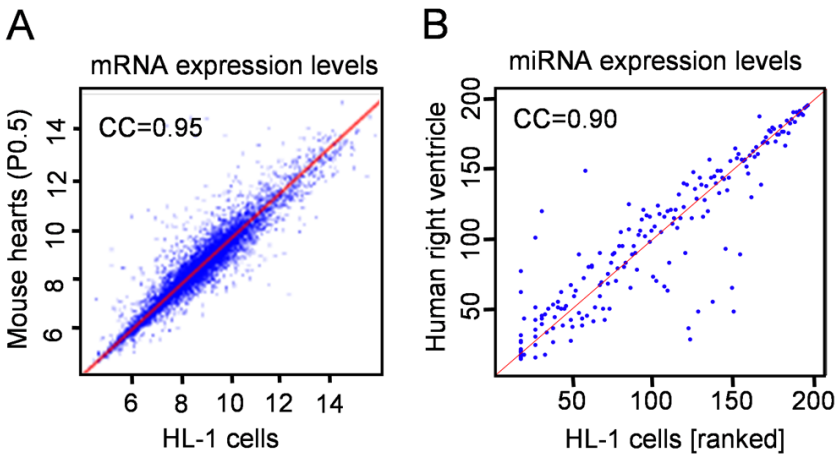

Figure 2.3: HL-1 mRNA and miRNA expression profiles are highly comparable to the ones observed in human and mouse hearts. (A) Gene expression levels obtained from HL-1 cells and P0.5 of C57/BL6 mouse heart. (B) Rank-transformed miRNA expression levels in HL-1 cells and human right ventricle.

To confirm and further investigate results from the analysis of the ChIP-chip data, additional ChIP-seq experiments were performed in this study, again using HL-1 cardiomyocytes now measuring Srf binding and H3ac sites on a genome-wide scale. Sample preparation was performed according the Illumina library preparation procedure. Two independent ChIP samples were profiled. After ChIP, DNA fragments bound by Srf or modified with H3ac in HL-1 cells were sequenced using the next-generation sequencing technology of the Illumina Genome Analyzer with short single-end reads of 36 bp in length. Sequencing was performed in-house at the Max Planck Institute for Molecular Genetics according to manufacturers' protocols. Analysis of the resulting images and successive base calling was done using the open source Firecrest and Bustard applications. Finally, deep sequencing of the ChIP libraries resulted in 6,967,318 and 8,364,328 reads obtained in the Srf and H3ac ChIP-seq experiment, respectively. The corresponding datasets have been analyzed in this study (see Chapter 4.2).

2.2.2 MicroRNA-seq after Srf Knockdown in Cell Culture

Considering that only a small proportion of differentially expressed genes in loss-of-function experiments are direct targets of the respective transcription factors, we studied the potential impact of miRNAs as secondary effectors (see Chapter 4.3). Again

	Srf siRNA-1	Srf siRNA-2	siNon
Total number of reads	14,911,499	14,518,157	14,742,382
Non-redundant read sequences	5,634,650	5,503,661	5,674,429

Table 2.1: Deep sequencing results of small RNA libraries of RNAi mediated knockdown of Srf (Srf siRNA-1/2) and non-specific siRNA (siNon) in HL-1 cardiomyocytes.

we focused on the transcription factor Srf, which is known to regulate cardiac-relevant miRNAs like miR-1 and miR-133[95,169]. To study if a significant reduction of the Srf protein in cardiomyocytes would affect the expression of associated miRNAs, a siRNA experiment was carried out using two siRNAs against Srf (Srf siRNA-1/2) and one non-specific siRNA (siNon) but now followed by miRNA quantification again using the NGS technology of the Illumina Genome Analyzer. Sequencing libraries were generated using a non-strand specific library construction method. Sequencing was performed in-house at the Max Planck Institute for Molecular Genetics according to manufacturers' protocols. Image analysis and base calling was performed using the open source Firecrest and Bustard applications. Deep sequencing of the small RNA libraries of RNAi mediated knockdown of Srf (Srf siRNA-1/2) and non-specific siRNA (siNon) control in HL-1 cardiomyocytes resulted in a huge amount of sequenced single-end reads of 36 bases in length, with much less unique (i.e. non-redundant) read sequences (Table 2.1).

2.2.3 MicroRNA-seq Data From Human Normal Heart

To evaluate MicroRazerS we used a dataset derived from three human normal heart (left ventricle) samples (see Chapter 4.3.1 for evaluation results). Small RNAs were isolated from total RNA using TRIzol (Invitrogen, Germany), pooled (3 times $\sim 3.5\mu$g total RNA was extracted and subsequently pooled) and prepared for Illumina GA sequencing according to the manufacturer's protocol. The sequencing library was generated using a non-strand specific library construction method. Sequencing was performed in-house at the Max Planck Institute for Molecular Genetics according to manufacturers' protocols. Image analysis and base calling was performed using the open source Firecrest and Bustard applications. Deep sequencing of the small RNA library produced 9,286,222 sequenced single-end reads of 36 bases in length, yielding 2,402,361 unique (i.e. non-redundant) read sequences.

2.2.4 RNA-seq, MicroRNA-seq and Genomic DNA-seq Data in Patients with Tetralogy of Fallot

In collaboration with the *German Heart Center Berlin* a broad panel of cardiac and blood samples from patients with congenital heart disease (CHD) as well as healthy individuals was collected. Each patient (sample) was phenotyped based on 250 anatomical and morphological characteristics. To identify key regulators in the cardiac development process and to investigate the interplay between different regulatory layers leading to CHD an integrative analysis of cardiac samples from patients with Tetralogy of Fallot (TOF), affected families and healthy unaffected individuals was performed. Syndromic cases and families with Mendelian inheritance were excluded. This analysis comprises the quantification of expressed mRNAs and miRNAs in patients with TOF as well as healthy individuals and targeted resequencing of a subset of cardiac samples and additional families with recurrent CHD. The results are given in Chapter 5.

mRNA and miRNA profiles were gathered from right ventricles of 22 patients with TOF as well as from left and right ventricle (LV and RV, respectively) of four healthy unaffected individuals (in total eight normal heart samples). The 22 cases of isolated TOF were selected out of a broad collection sampled in the German Heart Center Berlin, also balancing for age and gender (Figure 2.5). Total RNA was isolated using TRIzol (Invitrogen, Germany). mRNAs and miRNAs were isolated from total RNA and prepared for sequencing according to the manufacturer's protocol. Sequencing libraries were generated using a non-strand specific library construction method. Purified DNA fragments were used directly for cluster generation and 36 bp single-end read sequencing was performed using Illumina Genome Analyzer resulting in $\sim$19 million and $\sim$15 million reads per sample on average for mRNA and miRNA sequencing, respectively (Table 2.2).

Targeted resequencing was performed for 18 patients with TOF of which 13 are unrelated sporadic cases with very similar phenotype based on annotated disease characteristics and five are members of distinct families with recurrent CHD. Additionally, nine family members were sequenced consisting of seven healthy parents and two siblings affected with dextro-transposition of the great arteries (d-TGA) and tricuspid insufficiency (TI), respectively (pedigrees are shown in Figure 2.4). Genomic DNA (gDNA) was extracted from 14 out of 18 TOF patients as well as all family members from whole blood and for four TOF patients from right ventricle using standard protocols. The quality of gDNA was assessed on agarose gel and spectrophotometer. 3-5 μg of gDNA

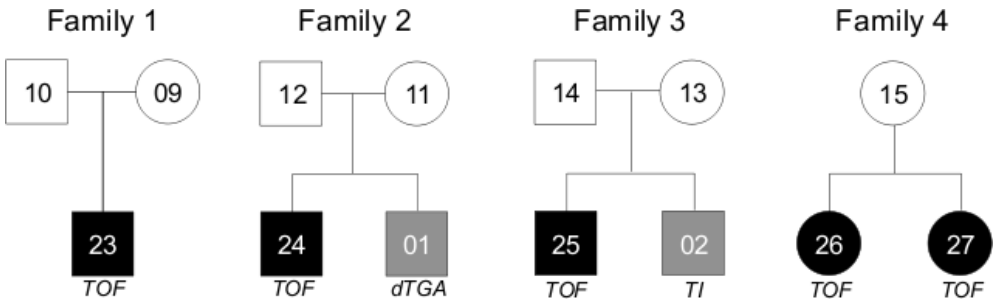

Figure 2.4: Pedigrees of four distinct families with recurrent congenital heart disease (CHD). Targeted sequencing of gDNA was performed for five Tetralogy of Fallot (TOF) patients and additionally for nine family members consisting of seven healthy parents and two siblings affected with dextro-transposition of the great arteries (d-TGA) and tricuspid insufficiency (TI), respectively. The numbers in the entities represent the sample identifiers (i.e. NH-{ID} for the healthy parents, TOF-{ID} and CHD-{ID}, respectively, for the affected children).

were used for Roche NimbleGen sequence capturing using 365K arrays. For resequencing we selected 867 heart- and muscle-relevant genes as well as 167 miRNAs based on knowledge gained in various related projects[1,170–172]. For sequence enrichment we applied NimbleGen sequence capturing using 365K arrays. For array design 12,910 exonic targets were selected representing 4,616,651 initial target bases, of which 97% (4,470,649 target bases) could be covered. DNA enriched after NimbleGen sequence capturing was pyrosequenced for 10 TOF patients using the 454 Genome Sequencer (GS) FLX instrument from Roche/454 Life Sciences using Titanium chemistry (∼430 bp reads), while the remaining samples were sequenced by Illumina Genome Analyzer (GA) IIx (36 bp paired-end reads). Sequencing was performed in-house at the Max Planck Institute for Molecular Genetics and by Atlas Biolabs (Berlin, Germany) according to manufacturers' protocols. The family samples were collected and prepared for target enrichment by the *Competence Network for Congenital Heart Defects* in Berlin. On average sequencing resulted in ∼13,271,000 read pairs and ∼759,000 single-end reads per sample for Illumina and Roche/454, respectively (Table 2.2).

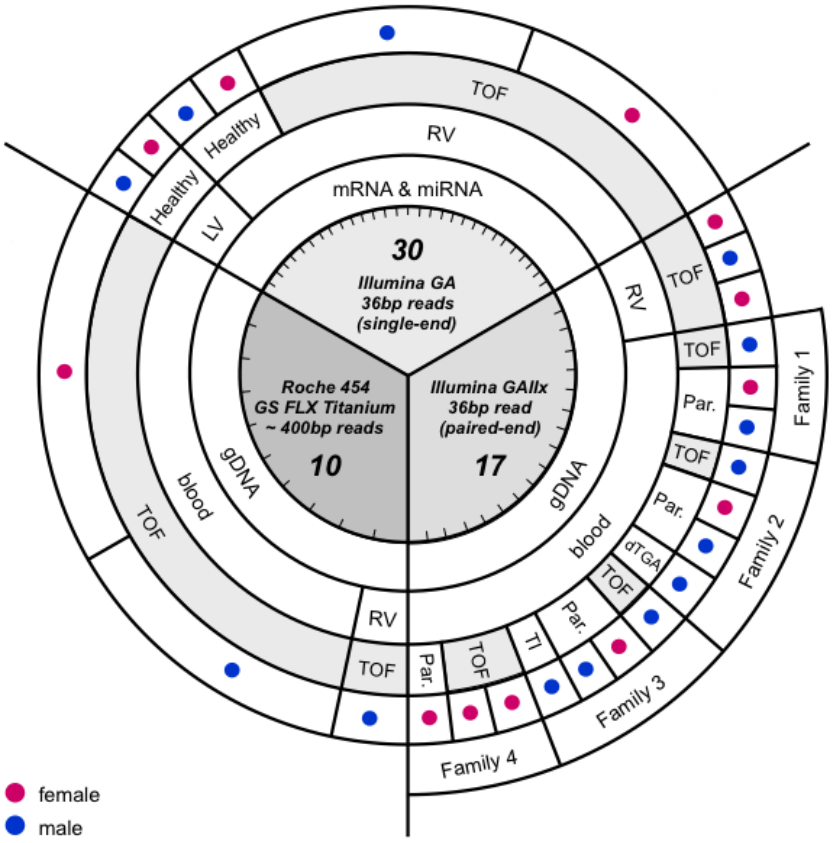

Figure 2.5: Overview about RNA-seq, miRNA-seq and gDNA-seq data in patients with Tetralogy of Fallot (TOF), affected families and healthy unaffected individuals (87 samples). Targeted resequencing was performed for 18 patients with TOF of which 13 are unrelated sporadic cases and five are members of distinct families with recurrent congenital heart disease. Additionally, nine family members were sequenced consisting of seven healthy parents and two siblings affected with dextro-transposition of the great arteries (d-TGA) and tricuspid insufficiency (TI), respectively. Cardiac samples were obtained from left and right ventricle (LV and RV, respectively), whereas most of the gDNA samples were obtained from blood. Next-generation sequencing was performed using different platforms including the Illumina Genome Analyzer (GA), the Genome Analyzer IIx (GAIIx) and the 454 Genome Sequencer (GS) FLX instrument from Roche/454 Life Science.

ID	Malformation	Family aggregation	Gender (Male/ Female)	Age category (years)	Source for lib prep	Single-end 36 bp read counts in mRNA-seq (GA)	Single-end 36 bp read counts in miRNA-seq (GA)	Paired-end 36 bp read counts in gDNA-seq (GAIIx)	Single-end 400 bp read counts in gDNA-seq (GS FLX)
NH-01	Normal	-	M	Adult (25)	LV	14,737,495	14,111,358	-	-
NH-02	Normal	-	M	Adult (25)	RV	20,106,950	16,270,049	-	-
NH-03	Normal	-	F	Adult (18)	LV	16,004,150	12,230,279	-	-
NH-04	Normal	-	F	Adult (18)	RV	12,961,101	12,940,172	-	-
NH-05	Normal	-	M	Adult (20)	LV	24,075,301	13,936,063	-	-
NH-06	Normal	-	M	Adult (20)	RV	20,296,818	14,475,968	-	-
NH-07	Normal	-	F	Adult (34)	LV	22,089,909	14,794,093	-	-
NH-08	Normal	-	F	Adult (34)	RV	23,597,799	14,890,970	-	-
NH-09	Normal	Fam 1	M	Adult	B	-	-	16,353,288	-
NH-10	Normal	Fam 1	F	Adult	B	-	-	11,512,571	-
NH-11	Normal	Fam 2	M	Adult	B	-	-	12,113,657	-
NH-12	Normal	Fam 2	F	Adult	B	-	-	15,134,005	-
NH-13	Normal	Fam 3	M	Adult	B	-	-	13,173,704	-
NH-14	Normal	Fam 3	F	Adult	B	-	-	12,178,506	-
NH-15	Normal	Fam 4	F	Adult	B	-	-	11,816,249	-
TOF-01	TOF	-	M	1-3 years	RV	10,888,508	15,618,489	15,971,391	-
TOF-02	TOF	-	F	1-3 years	RV	19,907,118	14,247,548	13,485,340	-
TOF-03	TOF	-	M	Infant	RV	21,882,581	16,154,319	-	-
TOF-04	TOF	-	M	Infant	RV	23,167,354	13,530,942	-	806,632
TOF-05	TOF	-	M	Infant	RV	14,570,039	13,178,983	-	-
TOF-06	TOF	-	F	1-3 years	RV/B	21,750,958	15,681,483	-	772,217
TOF-07	TOF	-	F	Infant	RV/B	18,392,413	14,459,386	-	833,654
TOF-08	TOF	-	F	Infant	RV/B	15,106,033	14,893,149	-	862,774
TOF-09	TOF	-	M	Infant	RV/B	23,512,940	16,226,821	-	744,316
TOF-10	TOF	-	F	Infant	RV/B	23,026,631	15,467,857	-	675,167
TOF-11	TOF	-	M	Infant	RV/B	17,430,948	14,989,342	-	850,429
TOF-12	TOF	-	F	Infant	RV/B	13,437,909	14,684,351	-	663,464
TOF-13	TOF	-	M	Infant	RV/B	21,026,718	15,412,115	-	663,583
TOF-14	TOF	-	M	Infant	RV/B	16,936,456	14,722,727	-	713,218
TOF-15	TOF	-	M	1-3years	RV	21,409,551	14,982,308	-	-
TOF-16	TOF	-	F	Infant	RV	16,813,107	16,914,098	-	-
TOF-17	TOF	-	F	Infant	RV	24,364,507	15,860,118	-	-
TOF-18	TOF	-	F	Infant	RV	20,193,649	16,542,142	12,738,154	-
TOF-19	TOF	-	M	Infant	RV	15,564,794	14,560,854	-	-
TOF-20	TOF	-	F	Infant	RV	21,553,557	17,891,078	-	-
TOF-21	TOF	-	M	Infant	RV	17,564,630	14,033,794	-	-
TOF-22	TOF	-	M	Infant	RV	24,353,916	16,296,019	-	-
TOF-23	TOF	Fam 1	M	Infant	B	-	-	10,442,596	-
TOF-24	TOF	Fam 2	M	Infant	B	-	-	12,741,583	-
TOF-25	TOF	Fam 3	M	Infant	B	-	-	15,275,837	-
TOF-26	TOF	Fam 4	F	Infant	B	-	-	13,939,375	-
TOF-27	TOF	Fam 4	F	Infant	B	-	-	12,059,011	-
CHD-01	d-TGA	Fam 2	M	Infant	B	-	-	12,380,168	-
CHD-02	TI	Fam 3	M	Infant	B	-	-	14,297,978	-

Table 2.2: Sample information and raw read counts obtained from RNA-seq, miRNA-seq and gDNA-seq in patients with Tetralogy of Fallot (TOF), affected families and healthy unaffected individuals. For library preparation, total RNA was isolated from left and right ventricle (LV and RV, respectively) of human heart samples and genomic DNA was obtained from blood (B) or RV. For sequencing different next-generation sequencing platforms were used including the Illumina Genome Analyzer (GA), the Illumina Genome Analyzer IIx (GAIIx) and the Genome Sequencer FLX (GS FLX) from Roche/454 Life Science.

Chapter 3

Computational Analysis of Next-Generation Sequencing Data

3.1 Mapping of Short Sequence Reads to a Reference Genome

Next-generation sequencing techniques support many applications including sequencing of chromatin-immunoprecipitated DNA for the identification of DNA binding sites and histone modification patterns, RNA sequencing for gene expression and small RNA profiling, and target resequencing for detection of genomic variations (Chapter 2.1). For all these applications, a vast amount of DNA is analyzed in terms of short sequences called reads, which represent fragments from a usually longer DNA molecule present in the sequencing sample. In contrast to whole-genome assembly, in which the sequence reads are assembled together to reconstruct a previously unknown genome, for these applications a reference genome is usually given[173]. One of the first computational challenges for analyzing the data of such applications is the mapping of all sequence reads to the reference genome. This read mapping problem can be formalized as follows: given a set of read sequences R, a reference sequence G and a distance $d \in \mathbb{N}$, find all substrings g of G that are within distance d to a read $r \in R$[174]. The occurrences of these substrings are called matches. Common distance measures are Hamming distance (mismatches and no InDels) and edit distance (mismatches and InDels)[174]. The mapping process is complicated by several factors including sequencing

errors, genetic variations in the population, short read length and the huge amount of reads to be mapped[175]. Therefore, many algorithms have been developed specifically for the purpose of mapping short reads (e.g. Bowtie[176], BWA[177], Eland[178], Maq[179], Novoalign[180], RazerS[174], SOAP2[181], SHRiMP[182] and ZOOM[183]).

The majority of the existing read mapping approaches use a filtration method followed by a verification step. The filtration method is first applied to identify candidate regions that possibly contain a match. In the following verification step these regions are examined for real matches. Often an index data structure, either on the set of reads or on the reference sequence, is build for filtration[174]. Several very successful filtering approaches use the q-gram counting strategy based on the q-gram lemma[184,185], which states that two sequences of length l with Hamming distance d share at least

$$t = l + 1 - (d + 1)q$$

common substrings of length q, so-called q-grams. This q-gram lemma can also be generalized to the edit distance if l is the length of the larger sequence[174]. Burkhardt and Kärkkäinen have described an extension that uses gapped q-grams[186]. The idea is to model insertions and deletions by additional q-grams. For example, with the basic shape 'N-N' applied the string, the pattern 'N-N', 'N--N' and 'NN' will be used. All three shapes in the pattern are compared to the q-grams of the basic shape in the string and therefore, matching q-grams can be found in the presence of InDels. The q-gram counting strategy was first used in QUASAR[187] and an improvement of this algorithm is the SWIFT filter algorithm[188], which relies on the q-gram filter for matches of error rate ϵ and a given minimum length l_0. Using an error rate rather than an absolute error threshold is more appropriate since the length of a local alignment is not known in advance.

Another algorithm which uses the q-gram filtering technique is SHRiMP[182]. However, the implemented default q-gram counting strategy in SHRiMP does not guarantee to be lossless. Therefore, Weese *et al.*[174] developed the short read mapping tool RazerS, which is implemented within the C++ library SeqAn[189]. It is also based on the q-gram counting strategy that builds an index over the reads and uses an implementation of the SWIFT filter algorithm to scan over the reference and efficiently filter regions containing possible read matches. These regions are identified by a certain minimal number t of q-grams. Filter efficiency is determined by the parameters q and t. For read-reference alignments that are not allowed to have gaps, i.e. if only Hamming distance mapping is

considered, filter sensitivity can be strongly increased by using gapped q-grams. RazerS can map sequence reads using Hamming or edit distance in the filtering phase and in the verification step without any restrictions. Moreover, given a user-defined loss rate (e.g. 0 making the mapping process exact), parameters are selected by the algorithm such that the chosen loss rate is not exceeded in expectation [174]. To map paired-end reads the reference genome is scanned from left to right in parallel with two SWIFT filters, which have the distance of the library (insert) size minus a tolerated deviation. Both filter search for potential matches of one of the two ends of all read pairs. In addition, all matches of the left filter within a distance of the doubled tolerated deviation are stored in a queue and if the right filter finds a potential match with corresponding stored by mate both potential matches are verified. To reduce the running time the verification process is only done if both potential matches are within the correct distance [174].

3.1.1 Small RNA Read Mapping Using MicroRazerS

Deep sequencing has become the method of choice for determining the small RNA content of a cell. Mapping the sequenced reads onto their reference genome serves as the basic for all further analyses, namely identification and quantification. Although specific short read mapping tools exists, several large-scale studies [190,191] have used the less sensitive and very time-consuming Mega BLAST algorithm [192] due to the special requirements of small RNA read mapping. Usually, a high quality 5' end with an exactly matching seed sequence and trailing mismatches at the 3' end is expected. As small RNAs may be shorter than the sequenced reads, the sequencing process can reach into the adapter. As a consequence, the 3' ends of the reads may contain variable lengths of adapter sequence causing mismatches in the read-to-reference alignment. If the adapter sequence is known, the 3' ends can be trimmed, but this process is imperfect and further complicated by the presence of sequencing errors occurring especially at the 3' end.

A promissing strategy for small RNA read mapping is therefore to search for the longest possible prefix-match of each read, i.e. the longest contiguous match starting at the first read base. Mega BLAST aligns all reads to the reference genome with a minimum word size. Its output needs to be further filtered for matches meeting the above criteria discarding all matches with lower than 100% identity in the 5' seed sequence and afterwards only retaining the longest match(es) for each read [190,191]. The resulting set of matches usually constitutes only a small fraction of the raw Mega BLAST output and moreover, this strategy is unnecessarily slow and inconvenient. However, there had

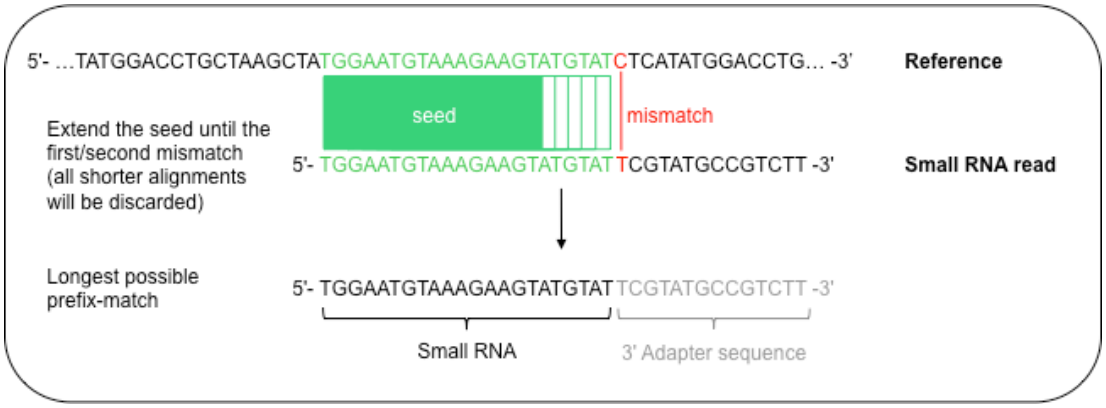

Figure 3.1: MicroRazerS strategy for the alignment of small RNA reads. The strategy is to search for the longest possible prefix-match of each read, i.e. the longest contiguous match starting at the first read base.

been no short read aligner that directly implements this strategy but there are tools employing similar strategies like the BWT-based aligners SOAP2[181] and Bowtie[176], which allow to set a minimum 5' seed length. Therefore, a read mapping tool, called *MicroRazerS*, specifically tailored to the needs of short RNA read mapping has been developed during this study[2]. MicroRazerS is robust to possible adapter sequence at the 3' end of a read and requires no adapter trimming. It can map millions of reads within a few minutes and is not only much easier to handle than Mega BLAST, but also more sensitive, especially in the presence of sequencing errors and SNPs. Moreover, no extensive filtering is required after mapping.

Like RazerS, MicroRazerS employs the gapped q-gram method in conjunction with the SWIFT parallelogram filter to detect with 100% sensitivity all read matches with a predefined read prefix of length s containing 0 or 1 mismatch. Seed matches are subsequently extended to the right (3' end) until the first mismatch is encountered. MicroRazerS thereby guarantees to find for each read the match that has (i) the lowest number of mismatches in the seed and (ii) can be extended furthest to the right. If multiple best matches exist, all of them are detected. The balance between speed and sensitivity can be controlled by the recognition rate parameter. The higher the recognition rate the more sensitive is MicroRazerS. The lower the recognition rate the faster runs the mapping tool (default 100). MicroRazerS supports seed length values from 10 to 26, a parameter that can be adjusted via the command line. If multiple best matches exist, a user-defined maximum number of hits is reported, optionally discarding all reads having more best hits than this number. An additional feature

of MicroRazerS is its option to map reads with at most one or no error in the seed sequence. Especially if one is interested in finding miRNAs at low abundance where robustness towards sequencing errors or SNPs might be crucial, the 100% identity criterium has to be dropped. A schematic representation of the MicroRazerS strategy for the alignment of small RNA reads onto a reference genome is shown Figure 3.1. Moreover, an evaluation of MicroRazerS in comparison to other short read mapping tools using similar strategies is described in Chapter 4.3.1.

3.2 Analysis of Protein-DNA Interactions from ChIP-seq Data

ChIP-seq has become the method of choise to investigate genome-wide in vivo binding patterns of transcription factors and chromatin histone marks. The analyses of protein-DNA interactions using ChIP-seq data is divided into (i) mapping of the obtained sequenced reads to the reference genome, (ii) normalization of read counts to account for experimental differences between different sequencing runs and (iii) calling of enriched sites (peaks). In this study ChIP-seq experiments using only a single sequencing run per experiment were performed. Therefore, no normalization of the resulting reads had to be performed. The read mapping is described in Chapter 3.1, and the peak calling with the corresponding discovery of sequence binding motifs is described in the following.

3.2.1 Peak Calling

After read mapping to the reference genome, the next step is to identify regions that are significantly more enriched than what would be expected by chance. For this task many peak calling algorithms have been developed mostly based on a sliding window approach. If a window of a given size contains a number of reads that exceeds a defined significance threshold, then this region is called a peak. There are algorithms that determine the background distribution (noise) from a control sample if available [193–197] while others model the background distribution from the ChIP sample itself [132,198]. Furthermore, a number of algorithms use the directionality of the reads, taking advantage of the fact that DNA fragments from a ChIP experiment are sequenced from the 5' end. The location of mapped reads should therefore form two peaks, one on the posi-

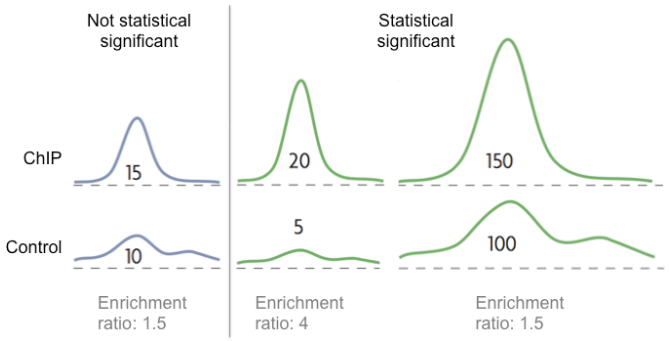

Figure 3.2: ChIP-seq peak scoring. The left ChIP peak is not statistically significant because the enrichment ratio between the ChIP and control sample is low (1.5) and the number of read counts (shown under the peak curves) is also low. The middle and the right peaks represent two ways in which a peak can be statistically significant. In the middle, the enrichment ratio between between the ChIP and control sample is high, although the number of read counts is low. On the right, the peaks have the same enrichment ratio as those on the left but have a larger number of read counts. Figure taken from Park *et al.*[134] and modified.

tive strand and one on the negative strand, with a constant distance between them[134]. Either by shifting each distribution towards the centre or by extending each region into an appropriately oriented fragment and then adding the fragments together a smoothed profile of each strand is constructed and the combined profile is computed[134]. This approach is used either to increase the statistical power of the peak detection[194,196,198] or to reduce the number of false positive peaks subsequently[193].

Based on the combined profile, a simple way to score a peak is a fold ratio of the reads from the ChIP sample relative to those of the control sample around the peak. This approach provides important information but nevertheless it is statistically not sufficient (Figure 3.2). Thus, the Poisson distribution has frequently been used to derive significantly enriched windows[194,195]. In addition, it can also be modified to account for regional biases in the read density due to chromatin structure, CNV or amplification bias[132,140,196]. However, Ji *et al.* have shown that the Poisson distribution does not perform well to model the background variability in real data[193]. They showed that a negative binomial (NB) distribution is much better suited than a Poisson distribution to model background distribution in the absence of a control sample by modeling both distributions on ChIP-seq data from mouse embryonic stems cells and comparing it to the observed control data. For peak calling, they used a sliding window approach to count the number of reads n in all non-overlapping windows of length w over the

genome. The Poisson distribution defines the probability of finding a number of k reads mapped to the window as

$$Pr(n = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Using a fixed rate λ Poisson model assumes that background reads are uniform distributed across all genomic loci. Ji *et al.* showed that this assumption does not fit well with the real data. Thus, they defined λ itself to be a random variable by assuming λ of window i to be gamma distributed

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

where $\Gamma(\alpha)$ is the gamma function. If α is a positive integer $\Gamma(\alpha) = (\alpha - 1)!$. For positive integer values as in ChIP-seq count data, exchanging the constant λ with $\gamma(\lambda)$ is equal to the negative binomial distribution $NB(\alpha, \beta)$ with probability

$$Pr(n = k) = \binom{k+\alpha-1}{\alpha-1} (\frac{\beta}{\beta+1})^\alpha (\frac{\beta}{\beta+1})^\beta,$$

where α and β are estimated to define the background distribution using counts for windows containing no or only a very low number of reads[134]. The observed number that a window contains k reads is compared with the expected number according to a null model. The ratio between the two numbers is used to calculate the false discovery rate (FDR) which is dependent on the window size. For peak calling a user-defined maximum FDR is chosen as cutoff determining a minimal read count per window. All windows that have a read count that exceeds this threshold are called enriched.

Difficulties in the identification of enriched regions are the different peak types including sharp and broad peaks (Figure 3.2). In general, sharp peaks are found for TFBS or histone modifications at regulatory elements, whereas broad peaks are often associated with histone modifications that mark domains such as transcribed or repressed regions[134]. The algorithm by Ji *et al.* implemented in the CisGenome software has been designed to handle both types of peaks by different sliding window approaches. In detail, a negative binomial distribution is used as the background model to estimate false discovery rates and this used error model allows the definition of a minimal FDR. CisGenome scans the reference genome with a sliding window of specified length

and identifies regions with a read count greater than a user-defined cut-off that do not exceed a specified FDR. Overlapping windows are subsequently merged into peaks. Moreover, CisGenome includes optional post-processing steps to enhance the peak detection. To obtain precise peak localization, localization boundary refinement can be applied. Reads coming from the forward and reverse strand are separated and the maxima of the individual strand-specific peaks are used to predict better boundaries for the enriched sites. Moreover, single-strand filtering can be applied which removes 5' without corresponding 3' peaks or vice versa[134].

3.2.2 Discovery of Sequence Binding Motifs

To determine binding characteristics and as a proof of principle, the analysis of protein-DNA binding experiments is often followed by a discovery of potentially causative binding sequence motifs. Based on the biochemical process of transcription factor binding to cis-regulatory elements in the promoter of their target genes, binding descriptors have been gathered for a large number of TFs[199,200]. The most common form to represent these motifs are position weight matrices (PWMs). PWMs represent motifs in a matrix form with one row per symbol of the alphabet $A = \{A, C, G, T\}$ and one column $i \in \{1, ..., L\}$ for each position in a pattern of length L. Each combination of symbol and position has a score assigned which typically represents the log-likelihood or, if a background nucleotide distribution is incorporated, the log-odds of observing that symbol at this position in the pattern. As a PWM assumes independence between positions in the pattern, the score between the PWM and the site with same length on the DNA sequence can be calculated as the sum of the individual symbol-position combinations. A common graphical representation for a PWM is the sequence logo[201]. In Figure 3.3 an example for a PWM, its sequence logo and real DNA binding site is given. PWMs can be used to predict the binding of a TF to the promoter sequence of their target genes. Two different approaches have been suggested. The more common approach uses predefined score cutoffs or PWM-derived statistics to predict individual binding sites for the TF. Examples are the MATCH program[203], the matching algorithm proposed by Rahmann *et al.*[204] provided by TRANSFAC[199] or the matrix-scan program[205]. The second approach biophysically models the binding of a TF to the full promoter sequence and predicts an affinity score which can be used to find likely bound promoters for each TF. This approach has been implemented in the TRAP algorithm by Roider *et al.*[206].

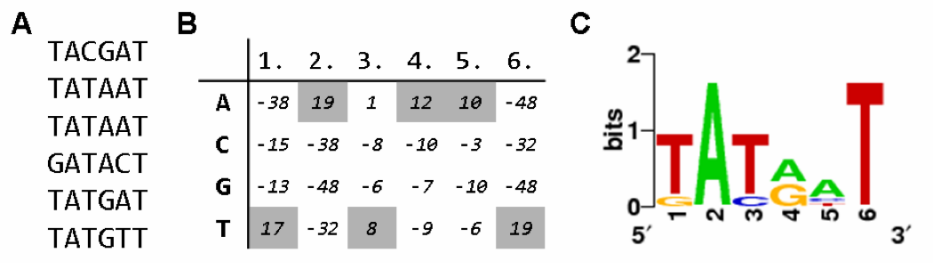

Figure 3.3: Different representations of a cis-regulatory element. (A) An example of six sequences corresponding to the -10 region of E. coli promoters. (B) The position weight matrix (PWM) of the same region using a large number of sequences. The best scoring nucleotide in each position is colored in gray and corresponds to the consensus sequence (TATAAT). (C) The sequence logo. The example is taken from Bulyk *et al.*[202].

The main drawback of computational approaches is the low signal-to-noise ratio which is commonly present in promoters of genes and leads to many false positive predictions. This problem is further aggravated by large distance between the actual binding site and the TSS. A common way to increase the signal-to-noise ratio is the use of information based on sequence conservation, e.g. obtained from alignments between the sequence of interest and an orthologous sequence from one or multiple related species. The idea is that regions with a strong regulatory impact are positively selected against mutations and therefore regions that show high variability can be discarded from the prediction of functional binding sites.

3.3 mRNA and Small RNA Profiling

RNA-seq is rapidly becoming the standard method for transcriptome analysis. A sensitive and accurate identification and quantification of known mRNA and miRNAs from mRNA-seq and small RNA-seq, respectively, is a key challenge to many of the applications of RNA-seq. The handling of sequenced reads that map to multiple genes or isoforms is an exemplary problem in the gene quantification, which is described in the following.

3.3.1 Quantification of mRNA Expression Levels

To measure a gene's (g) expression level by NGS reads (r), an obvious way is to determine its read count c(g,r), which is the number of reads mapping to the set E of all its exons $e_1, ..., e_n$. Appropriated gene model can be derived from databases such as ENSEMBL[207] or RefSeq[208].

$$c(g, r) = \sum_{e \in g} c(e, r_{exonic}) + \sum_{e \in g} c(e, r_{junction}),$$

where r_{exonic} is the number of reads that are fully included in exons, called exonic reads, of protein-coding genes and $r_{junction}$ is the number of assigned junctions reads to an exon. Junction reads overlap two or more exons and are often assigned proportionately to each of their overlapping exons. For genes encoding multiple isoforms, the number of hits (i.e. exonic and junction reads) per gene is determined as the sum of all hits over all possible exons.

One of the main problems with mapping short reads is the significant number of reads that map to multiple positions in the reference genome, mostly attributed to paralogous genes, low complexity and repetitive sequences[147]. The fraction of these multi-mapping reads varies and depends on the transcriptome and read length. As an example, for the datasets analyzed by Li *et al.* this fraction ranged between 17% (mouse) and 52% (maize) for 25 bp reads, representing a significant proportion of RNA-seq data[209]. However, longer reads do not decrease the number of multi-mapping reads as much as expected. The simulations on mouse transcriptome in Li *et al.* showed that single-end and paired-end (200 bp insert) reads with length of 75 bp give rise to 10% and 8% multi-mapping reads, respectively[209].

There are different approaches in the handling of multi-mapping reads including keeping only uniquely mapped reads, mappability methods, rescue methods and statistical models. The most straightforward approach is to discard multi-mapping reads. This has been often done in the first RNA-seq studies[210,211]. Keeping only uniquely mapped reads can introduce experimental bias including an underestimated expression of repetitive genes. A more sophisticated method using only uniquely mapped reads adjusts the read count for each exon by its mappability, i.e. an essentially fraction of exon positions that give rise to uniquely mapping reads[145]. Consider a given genomic position i and let s_i be an n-mer subsequence that starts at this genomic position. Let P_i be the set of positions to which the n-mer s_i maps. If the n-mer is unique, its position set contains a single entry $P_i = \{\text{i}\}$. For multi-mapping positions of the n-mer $|P_i| > 1$.

Let $u_i = 1$ if $P_i = \{i\}$ and $u_i = 0$ otherwise. Let Q_i be the set of all genomic positions that neighbor on position i and start an n-mer that overlaps with genomic position i. The mappability m_i is defined for each i as the fraction

$$m_i = \frac{\sum_{j \in Q_i} u_j}{n},$$

which results in the number of unique mappable n-mers that overlap position i[145]. The mappability is one if each n-mer that overlaps with position i is unique in the reference genome. However, this mappability method also introduce experimental bias by discarding sequencing data although it corrects for repetitive sequence bias[145].

One strategy that uses all sequencing data is to rescue multi-mapping reads by allocating fractions of them to genes in proportion to coverage by uniquely mapping reads[209]. In the rescue method implemented in the ERANGE (Enhanced Read Analysis of Gene Expression) package multi-mapping reads are assigned fractionally to their different possible locations based on using the calculated initial expression levels from the unique reads of their respective gene models[147]. This rescue method has been implemented for gene-level expression only. Another rescue method, shown to be not as sensitive to errors in gene annotation, is based on a local window approach. In the MuMRescue approach multi-mapping reads are proportionately assigned to each of their mapping locations based on unique coincidences with uniquely mapped and other multi-mapping reads[212,213]. This is archived by counting the uniquely mapped reads that occur in a specific window around each locus occupied by a multi-mapping read divided by the total number of uniquely mapped reads proximal to genomic locations associated with that multi-mapping read[213]. Both rescue strategies have been shown to improve correlation with microarray data[147].

Among reads that map to multiple positions in a reference genome, it is also possible that reads map to a single gene but multiple isoforms, called isoform multi-mapping reads. A method that handles isoform multi-mapping reads by explicitly estimating isoform expression levels but not handling gene multi-mapping reads was published by Jiang and Wong[214]. They used the Poisson distribution and the maximum likelihood estimation via coordinate-wise hill climbing to determine isoform expression levels. The individual parameters are optimized until convergence and confidence intervals are estimated using an importance sampling approach[214]. Finally, a statistical model has recently been suggested to estimate individual isoform expression levels and more-

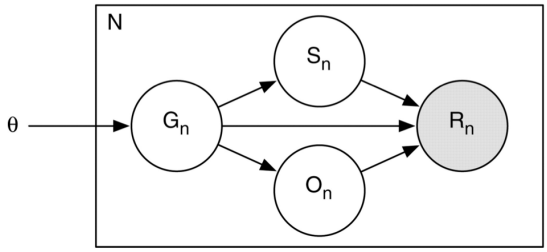

Figure 3.4: Graphical model for RNA-seq data. Figure taken from Li *et al.*[209].

over, incorporate gene multi-mapping reads[209]. Interestingly, it has been shown that previous rescue methods are approximately equivalent to one iteration of Expectation-Maximization (EM) algorithm[209]. A Bayesian network (Figure 3.4) is used to estimate gene and isoform expression levels. The model generates N independent identically distributed reads of length L. The sequence reads (observed data) are represented by the R_n random variables and each read is associated with three hidden random variables, G_n (isoform), S_n (start position) and O_n (orientation) from which the read was derived. The primary parameters of the model $\Theta = [\Theta_0, ..., \Theta_M]$ correspond to the expression levels, assuming that all M isoforms present in the transcriptome are given. The full data likelihood for this model is

$$P(g, s, o, r|\Theta) = \prod_{n=1}^{N} P(g_n|\Theta)P(s_n|g_n)P(o_n|g_n)P(r_n|g_n)P(r_n|g_n, s_n, o_n).^{209}$$

The random variable G_n takes a value from 0 to M, with 0 representing noise, i.e. reads that do not map to known transcripts. The random variable S_n takes a values from 1 to $max_i l_i$, where l_i is the length of isoform i. The random variable O_n is binary and indicates if a read is in the same orientation as the parent isoform or the reverse complement. The hidden random variables for the n-th reads can be summarized with a set of indicator random variables Z_{nijk}, where $Z_{nijk} = 1$ if $(G_n, S_n, O_n) = (i, j, k)$. For strand-specific protocols the variables $Z_{nij} = Z_{nij0}$ are used. To find the maximum likelihood values for Θ the EM algorithm is used. In general, in the expectation (E) step the expected values of Z_{nijk} random variables, given the current parameter values Θ, are computed. For a strand-specific protocol and a uniform read start position distribution (assuming that reads are generated uniformly across isoforms), this computation is

$$E_{Z|r,\Theta^t} = \frac{(\Theta_i^t/l_i)P(r_n|Z_{nij}=1)}{\sum_{i'j'}(\Theta_{i'}^t/l_i)P(r_n|Z_{ni'j'}=1)}.^{209}$$

After computing the expected read counts in the E-step, the following maximization (M) step computes expression values maximizing likelihood given expected read counts. The parameter-estimates are then used to determine the distribution of the hidden variables in the next E-step. Both steps, the E- and M-step, respectively, are repeated until convergence. The model estimates maximum likelihood expression levels using the EM algorithm[215].

3.3.1.1 Isoform Quantification using POEM

In this study a proportion estimation (POEM) method[21] that enables the relative quantification of known isoforms using model assumptions similar to those of Jiang and Wong[214] was used and in addition, optimized for analyzing the RNA-seq datasets described in Chapter 2.2.4. The POEM method is implemented in *Solas*, a package for the statistical language *R*. In general, the algorithm was designed to estimate the abundance of each known isoform based on a probabilistic model that integrates the number of reads in exons and the information pertaining to annotated isoforms such as the sequence read mappability of their related exons[21]. The total number of reads R covering an isoform j is determined by a Poisson process

$$R_j \sim Poisson(\lambda \cdot s_j \cdot p_j),$$

where s is the total length of the isoform, p is the relative proportion and λ is a normalizing factor related to the sampling depth. Especially for low-coverage datasets the Poisson model serves as a bettter approximation than the normal distribution[216]. Moreover, this distribution has already been proposed for abundance of expressed sequence tags (EST data)[217] and SAGE libraries[218]. To infer the non-observed proportions p_j of the isoforms again the EM algorithm is used.

The analysis of alternative splicing events showed that frequent splicing events are occurring on the most 3'- or 5'- exons and therefore, the first and last exon of every transcript is artificially removed before POEM estimation[21]. Moreover, due to different 3'UTRs or alternative exons lengths, there are overlapping exons between the different isoforms of a gene. For correct POEM estimation, these exons should be removed in order to get only regions which non-ambiguously describe every isoform. However, removing overlapping exons results in under- or overestimation of specific isoforms de-

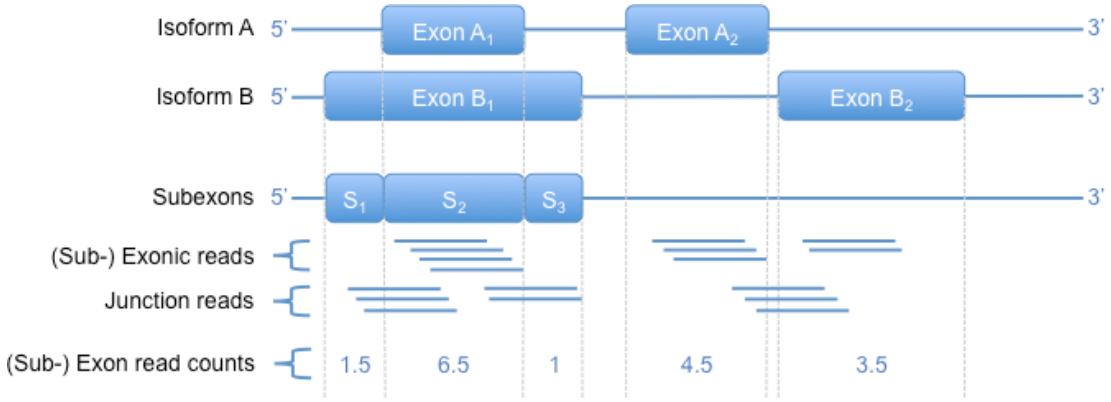

Figure 3.5: Modified gene model for isoform estimation using the POEM method. Example of a gene with 4 exons and 2 isoforms. All overlapping exons are cut down into subexons, i.e. the two overlapping exons A_1 and B_1 produce three subexons S_1, S_2, and S_3 for the gene model. For the final read count of an exon or subexon the (sub-)exonic and junction reads are counted. Subexonic and junction reads are proportionately assigned to each of their overlapping (sub-)exons.

pending on their exonic read counts.

To apply the POEM method to a gene, two information have to be specified including (i) the description of the gene model (i.e. exon coordinates and isoform structures) and (ii) the read counts observed within the exons. To optimize the estimation both information are modified in this study to keep overlapping exons and moreover, to integrate splice junction counts for the estimation of isoform proportions, which is also missing in the original POEM estimation. In a first step, all overlapping exons are cut down into subexons. For instance, two exons partially overlapping should produce three subexons for the gene model (Figure 3.5). In a second step, the read counts observed within the exons and subexons are (re-)defined. For all non-overlapping exons the number of reads that are fully included in the exon boundaries are counted as described above. In addition, junction reads that overlap two or more exons are also included in the read count of an exon in the way that these are proportionately assigned to each of the overlapped exons. For the subexons, the number of reads that are overlapping their boundaries by at least one base are counted by proportional assignment to all overlapped subexons. This counting approach includes subexonic reads that are fully included in the subexon boundaries as well as their junction reads that overlap two or more subexons. Finally, the read counts observed within the exons comprises reads of exons and overlapping exons as well as junction reads between exon as well as between subexons. However,

the model can be extended to include junction reads in more probabilistic way instead of adjusting just the corresponding exonic read counts.

3.3.2 Quantification of MicroRNA Expression Levels

After mapping small RNA reads to the reference genome the genomic mapping information of each read are used for small RNA annotation. Reads are annotated based on their overlap to known genome annotations including miRNAs, other non-coding RNAs, repeating elements and protein-coding regions. Annotations are obtained from UCSC database (GenBank mRNA, RepeatMasker and sno/miRNA tracks)[219] and miRBase (miRNAs)[43].

If a read overlaps to a known mature miRNA sequence (or known precursor hairpin sequence) in the correct orientation, then it is assumed to be a sequencing product of this miRNA and is added to its read count. Multi-mapping reads are proportionally assigned to each of their loci or miRNAs.

A typical small RNA-seq sample consists of a number of other non-coding RNAs besides miRNAs including transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small cytoplasmic RNAs (scRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), miscellaneous RNAs (miscRNAs), mitochondrial tRNA-derived pseudogenes (mt-tRNAs) and 5S ribosomal RNAs (5S rRNAs). Commonly, the most abundant classes of small RNAs besides miRNAs in a given sample are rRNAs and tRNAs. The 5S rRNA is commonly used for normalization in miRNA qRT-PCR experiments[220–222]. However, Peltier *et al.*[220] showed that these and other commonly used reference RNAs used in miRNA qRT-PCR experiments, such as 5S rRNA, U6 snRNA or total RNA were the least stable against the most consistently expressed miRNAs across 13 discrete normal human tissues. Their data suggests that total RNA is inferior to the most consistently expressed miRNAs and 5S or U6 were the two least stable RNA species. The standard deviation across all tissue samples when normalized to 5S rRNA was the highest followed by those from U6 snRNA[220]. Besides small non-coding RNAs reads are further mapping to different classes of repeating elements including short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), long terminal repeat elements (LTR), DNA repeat elements, simple repeats (micro-satellites), low complexity repeats and satellite repeats. In mammals, the most common elements are LINEs and SINEs (including ALUs).

The different kinds of small RNA annotations can be used as an additional information to evaluate the quality of the underlying small RNA library preparation for NGS. For example, a low number of reads that corresponds to small non-coding RNAs except miRNAs indicates an accurate library preparation and a low number of mRNA reads points to a low contamination of the total small RNA sample.

3.4 Differential Expression Analysis

3.4.1 Quality Control

Before normalization and computation of differential expression, the samples should be examined to identify possible outliers. Eliminating the impact of outliers can significantly improve the precision of normalization and a good quality control analysis can cope with technical artifacts and variance in the experiments.

Among others the principle component analysis (PCA) can be used to examine the global patterns across samples. PCA is a statistical technique for exploring the structure of high dimensional data, such as those generated from NGS experiments. In simple terms PCA reducing data dimensionality by searching for dimensions with highest variance and subsequent projection which allows to visualize sample relationships in the context of experimental factors. Thus, factors can be inferred which are the key to the variances in the observations (e.g. gene or miRNA expression)[223]. Potential inferences can be drawn according to e.g. the library preparation and contamination of the RNA-seq libraries. Another common dimension reduction method is multi-dimensional scaling (MDS). While PCA finds linear combinations of the variables to get the most variation in multivariate data, MDS aims to preserve proximity and distance between pairs of cases. Classical MDS is identical to PCA for most datasets, however, if one dimension is fixed, the samples can be place in an arrangement that is often more representative of true distances[224].

In RNA-seq experiments library preparation and sequencing can introduce systematic biases and artefacts like over-amplification of GC-rich regions and generation of duplicate sequences. Unfortunately, it is difficult to distinguish between reads that represent potential PCR artefacts and normal duplicate reads. It might be that stacks of exactly duplicated reads (pile-ups) indicate mapping or PCR problems, or they could reflect a true signal. Thereby, removing all duplicate reads might causes underestimation of

the true (real) read count level. However, in particular cases, duplicated reads must be removed, an example being the detection of SNPs, fusion transcripts or to get the real depth of coverage for a genomic region. In a diverse sequencing library most sequences are expected to occur only once in the dataset. A low level of duplication may therefore indicate a very high level of coverage of the target sequence, yet a high level of duplication is more likely to indicate some kind of enrichment bias such as PCR over-amplification. In summary, the level of duplication in a sequencing library should be examined individually and also in comparison to other libraries, potentially resulting in resequencing of the library.

3.4.2 Normalization

For accurate estimation and detection of differential expression, normalization is a critical step which aims to remove any systemic technical effects that might occur in the data to ensure that technical bias has as low impact as possible on the results. In RNA-seq experiments, RNA systematic technical bias originated by the reverse transcription reaction, RNA ligase preferences and PCR based amplification during library preparation are frequent as well as composition bias due to relying on library size[225,226]. Small RNA-seq experiments are strongly biased towards certain small RNAs largely independent of the sequencing platform but strongly determined by small RNA library preparation method[226].

To normalize data between samples typically the total number of reads in a given lane or library is scaled to a common value across all sequenced libraries in the experiment. For example, in many approaches the observed counts for a gene are modeled by the mean and an additional factor modeling the total number of reads in the library[211,227,228]. For LongSAGE-seq data, the square root of scaled counts[229] or the beta-binomial model[230] is used, both using the total number of observed read counts[225]. Mortazavi *et al.* adjust the counts to reads per KB per million mapped reads (RPKM), defined as

$$RPKM = 10^9 \frac{C}{NL},$$

where C is the number of mappable reads that fell onto the gene's exons, N is the total number of mappable reads in the experiment and L is the sum of the exons in base pairs[147]. By contrast, Cloonan *et al.*[148] log-transform the gene length-normalized read

count data and apply quantile normalization and moderated t-statistics as in microarray normalization[225]. Sultan *et al.*[231] normalize read counts by the virtual length of the gene, the number of unique k-mers in exonic sequence as well as by the total number of sequenced reads[225]. Bullard *et al.* used an upper-quartile normalization method, in which counts are divided by upper-quartile of counts for transcripts with at least one read[232].

For small RNA-seq experiments library size scaling is a common procedure for normalization. Following this method the reads assigned to a miRNA (or small RNA) are divided by the total number of small RNA-seq reads mapped to the reference genome[233]. Alternatively, the relative frequency of miRNAs is determined by normalizing miRNA reads against the total number of reads that mapped to known miRNAs[226,234]. However, this normalization approach has its limitations for datasets with markedly different RNA compositions which could be affect this number[225,233]. Therefore, Robinson and Oshlack suggested the trimmed mean of M-values (TMM) normalization method to remove RNA composition bias. They argue that the number of reads for a RNA or small RNA is dependent not only on its expression level and length, but also on the RNA population from which it originates[225]. For the sample framework Robinson and Oshlack define Y_{gk} as the observed read count for gene g (or miRNA) in library k, μ_{gk} as the true but unknown expression level, L_g as the length of g and N_k as total number of reads for library k. Then they model the expected value of Y_{gk} as

$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k,$$

$$\text{where } S_k = \sum_{g=1}^{G} \mu_{gk} L_g.^{225}$$

The total RNA output of a sample is represented by S_k. While N_k is known, S_k is unknown and can vary widely from sample to sample, depending on the RNA composition. If a RNA population has a larger total output, then RNA-seq experiments will under-sample e.g. miRNAs or mRNAs, relative to another sample[225]. Since the expression levels and the true length of every gene is unknown, S_k cannot be estimated directly. However, the relative RNA population of two samples $f_k = S_k / S'_k$ can be estimated by using a weighted trimmed mean of the log expression values. For sequencing data, Robinson and Oshlack define the gene-wise log-fold changes as

$$M_g = log_2 \frac{Y_{gk}/N_k}{Y_{gk'}/N_{k'}}$$

and the absolute expression levels as

$$A_g = \frac{1}{2}\ log_2(Y_{gk}/N_k\ \bullet\ Y_{gk'}/N_{k'}) \text{ for } Y_{g\bullet} \neq 0.$$ [225]

Both the M values and the A values are trimmed before taking the weighted average. The TMM method assumes that the majority of genes or small RNAs common to both samples, are not differentially expressed. Conducted simulation studies have shown that the method is robust against deviations to this assumption up to approximately 30% of differential expression in one direction[225,235]. The Bioconductor package *edgeR*[236] comprises e.g. RNA composition adjustment by TMM and quantile-to-quantile count adjustment. This approach is used to adjust the observed counts up or down depending on whether the corresponding library sizes are below or above the geometric mean (called qCML for quantile adjusted conditional maximum likelihood) which creates approximately identically distributed read counts (pseudodata).

3.4.3 Defining Differential Expression

Early methods for differential expression between two or more sequencing libraries pooled the libraries in each class and used a standard two-sample difference in proportions test or Fisher's exact test[237]. Yet, this pooling deals inadequately with the within-class variability[238–240]. Moreover, for each class (i.e. consider a two-sample comparison, e.g. patients versus healthy individuals) the number of pooled libraries must be equal. A more flexible model computed two-sample t-statistics on the proportions[241], thereby taking into account the library-to-library variability. More natural choices for a statistical model of tag counts may be Poisson or Binomial. However, in practice there are library-to-library variations which are not well captured by these distributions. The mean-variance relationship of either Poisson (assuming that the mean is equal to the variance) or Binomial distribution may not provide enough flexibility, i.e. more variability exists than can be explained by the model (this is called overdispersion). A better fit therefore requires the specification of extra model parameters. More recent methods have explored the use of beta-binomial[238,239] and gamma-Poisson (negative binomial)[240] models. Lu *et al.* showed via simulation studies that the negative binomial model seem to performs superior[240].

There are several R packages available from Bioconductor that allow to analyze differential expression in digital gene (or miRNA) expression datasets. These include *edgeR* and *DESeq*, which use an exact test based on NB distribution[236,242], *DEGseq* which implements MA-plots using random sampling model or technical replicates and assumes normal distribution of M given A[243–245], and *baySeq* which uses an estimation of the posterior likelihood of differential expression via empirical Bayesian methods based on Poisson or NB distributions[246]. The main differences between each package is how the dispersion (or variance) is calculated.

In this study the gene and miRNA expression datasets have been analyzed in respect to differential expression using the edgeR implementation based on a negative binomial model for count data. It states

$$Y_{ij} \sim NB(\mu_{ij}, \phi)$$

$$\text{with } E(Y_{ij}) = \mu_{ij} \text{ and } Var(Y_{ij}) = \phi + \phi\mu_{ij}^2,$$

where ϕ is the dispersion (for $\phi = 0$ this resembles the Poisson distribution) and Y_{ij} is the observed count for class i and library j for a particular tag. If λ_i is the true relative abundance of this tag in RNA of class i then $\mu_{ij} = m_{ij}\lambda_i$ where m_{ij} is the library size for sample j. Differences in relative abundance are assessed for each tag by testing the null hypothesis $H_0 : \lambda_1 = \lambda_2$ against the two-sided alternative $H_1 : \lambda_1 \neq \lambda_2$. In detail, there are two alternatives. First, assuming that all tags have the same dispersion, all tags are used for estimation (hard shrinkage), or second, the estimate of individual tag dispersions is modulated by sharing information among all tags (soft shrinkage or weighted likelihood)[237].

In most methods the inference is done one-tag-at-a-time, which is equivalent to gene-wise t-test for differential expression in microarray studies. In the extreme case two libraries vs. one, one-tag-at-a-time inference would require the estimation of three parameters from three observations. Moreover, Robinson and Smyth have observed that the estimation of the overdispersion ϕ can be problematic, especially in very small samples[237]. Therefore, they share information over all tags to improve the inference using a procedure analogous to the empirical Bayesian method implemented in the *limma* package. As a result, the standard t-statistic is replaced with a moderated t-statistic[247].

For SAGE data Robinson and Smyth discuss a common dispersion model, which uses all tags to estimate a common dispersion. The conditional likelihood for a single tag is formed by conditioning on the sum of counts for each class, where the sum of identically distributed NB random variables also follows NB. However, in the frequent situation of unequal library sizes, the counts are not identically distributed. Therefore, they used qCML normalization which creates pseudodata that can be inserted into the equation for the single-tag conditional log-likelihood for ϕ, summed over all tags and maximized with respect to ϕ, resulting in a common estimate. For statistical testing the difference in expression between two conditions like patients versus healthy individuals they used the above described exact test [237]. The assumption of a common dispersion offers a significant stabilization, compared with a tag-wise estimation, especially for very small samples. However, in reality not each tag has the same dispersion, implying that inference can be improved by a less strong stabilization. Therefore, instead of enforcing a common dispersion on all tags, they proposed to squeeze each tag-wise dispersion (i.e. individual estimate denoted as ϕ_g) towards common dispersion estimate (similar to empirical Bayesian). They define the weighted log (conditional) likelihood WL (ϕ_g) to be a weighted combination of the individual and common likelihoods as

$$WL(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g),$$

where α is the weight given to the common likelihood [237]. If $\alpha = 0$ this formula resembles the tag-wise qCML estimates, meaning that the common dispersion was sufficient. For $\alpha >> 1$, the contributions from any individual log-likelihood is outweighed by the common likelihood and the result is a common dispersion. If the true dispersion is quite variable, $\alpha \approx 1$, and if a large number of samples is given, sufficient individual estimates can be obtained. Improved dispersion estimation enhances inference of differential expression (i.e. requires an approximate level of squeezing α). One possibility is to select α tag-wise, as some tags may need more squeezing.

3.5 Correction for Multiple Testing

Analyzing large-scale biological data like involves the repeated performance of statistical tests. A p-value without correction for multiple testing is only statistically valid when a single score or a very low number of scores is computed. For example, if a

single gene had been tested to be differentially expressed between two conditions, the p-value could be used directly as a statistical confidence measure. However, performing the same test 10,000 times, one would expect $10,000 \cdot 0.01 = 100$ of them to have a p-value ≤ 0.01, even in a completely random situtation. Due to thousands of hypotheses that are tested simultaneously (multiplicity problem) the chance of false positives significantly increases. Therefore, we need to adjust for multiple testing based on the number of tests performed when assessing the statistical significance of analyses of high-throughput datasets.

To correct for the increase in false positives classical methods aims to ensure a least overall family-wise error rate by adjusting the individual hypothesis significance levels. The most widely used method of multiple testing correction is the Bonferroni adjustment, which distributes the significance threshold α evenly on all separately performed tests n by requiring a significance threshold of at least α/n. However, this method is too conservative, especially for the analysis of high-throughput data where the number of tests can easily exceed many thousands resulting in only very low numbers of significant tests[248]. For these kind of analysis methods that control the false discovery rate (FDR), which is the expected proportion of false discoveries among all significant tests, are more valid. The method to control the FDR in this study was originally introduced by Benjamini and Hochberg[249] for independent p-values and was later adapted by Benjamini and Yekutieli[250].

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

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

Subsequently, all the hypothesis tests with p-values less than or equal to P_k are rejected. The two methods differ 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. Therefore they proposed a more conservative estimations of the FDR

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

which does not require independency of the p-values[249,250]. Finally, Benjamini-Yekutieli FDR-adjusted p-values can be computed using a step-wise procedure, each representing the lowest level of FDR, where the appropriate hypothesis belongs to the set of rejected hypothesis for the first time[251,252].

3.6 MicroRNA Target Prediction

MicroRNAs are involved in the regulation of protein expression in plants and animals. Predominantly, they bind to the 3'UTR of mRNAs to inhibit translation or to induce cleavage. MicroRNAs can have hundreds of different targets in a cell and most miRNAs in plants show near perfect complementarity to their targets[253–255]. In animals miRNA-target prediction was shown to be more complex because only few miRNAs are perfectly complementary to their targets. Different computational methods have been developed for miRNA target prediction and in the following utilized prediction tools and their principles are presented.

3.6.1 Principles of Target Prediction

The probably most important factor for miRNA target prediction is the Watson-Crick paring to the 5' region of the miRNA centered on nucleotides 2-7, which is called the miRNA seed[256]. Requiring a Watson-Crick seed pairing substantially improves the performance of computational target prediction and reduces notably the occurrence of false positives. Most miRNA targets have only a single 7 nt match to that miRNA seed region. Either nucleotides 2-8 build base pairs (7mer-m8; Figure 3.6D) or nucleotides 2-7 build base pairs combined with an A across position 1 (7mer-A1; Figure 3.6C). The A-anchor across nucleotide 1 is shown to be conserved in vertebrates[257] and moreover, there is experimental evidence that the 7mer-A1 sites outperform others with a Watson-Crick match to position 1[258,259]. Requiring perfect 8 nt seed pairing (8mer; Figure 3.6E) increases specificity, whereas 6 nt pairing (6mer; Figure 3.6A-B) increases sensitivity. Thus, the site efficacy can be ranked as follows: 8mer >> 7mer-m8 >> 7mer-A1 >> 6mer > no site, with the 6mer differing only slightly from no site at all[256,259,260]. In addition to the 5' seed pairing, pairing to the 3' end of miRNAs also plays a role, although a minor ones, in target recognition[256]. The miRNA usually supplements seed pairing to improve binding specificity and affinity. Such 3'-supplementary pairing ideally centers on miRNA nucleotides 13-16 with at least 3-4 contiguous pairs and the UTR region

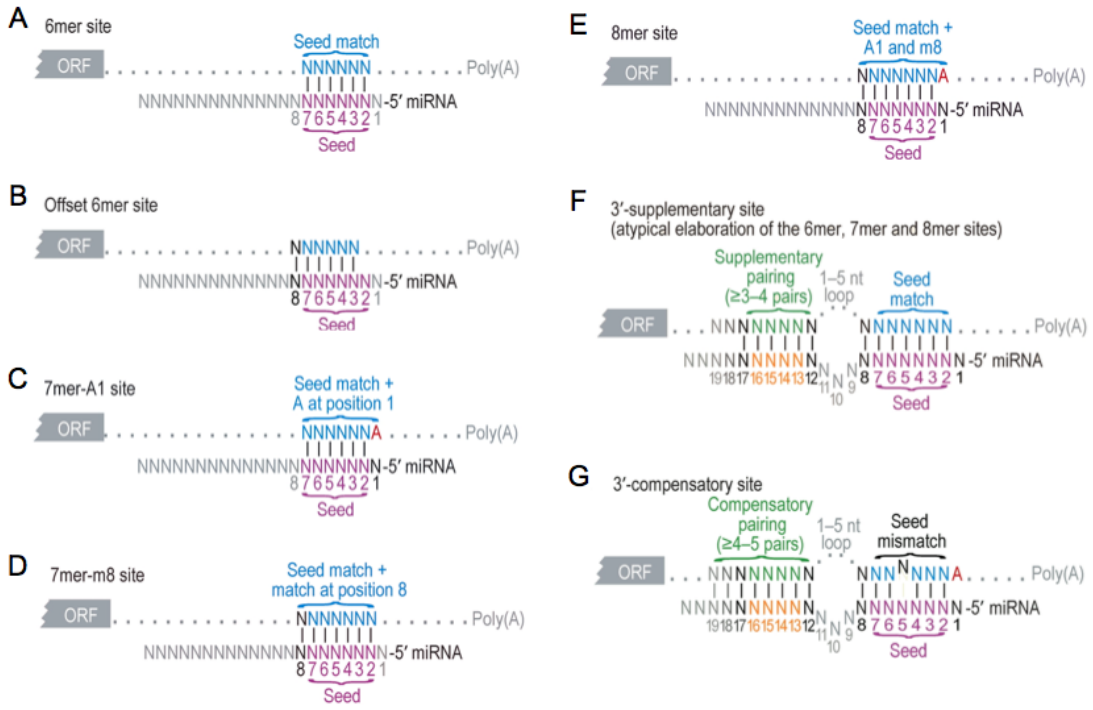

Figure 3.6: Different types of miRNA target sites. (A-B) Marginal, 6 nt sites matching the seed region. (C-E) Canonical, 7-8 nt sites matching the seed region. (F-G) Atypical sites with 3' supplementary and compensatory pairing, respectively. The pictures are taken from David Bartel[256].

directly opposite this segments (Figure 3.6F). However, supplementary 3' pairings are very rare and play a modest role in target recognition. Moreover, pairing to the 3' region of the miRNA can also compensate for a single nucleotide bulge or mismatch in the seed region[256]. These are called 3'-compensatory sites and the pairing centered on miRNA nucleotides 13-17 extends to at least nine consecutive Watson-Crick pairs (Figure 3.6G). 3'-compensatory sites are rare and probably emerge only when a specific member of a miRNA family is required for regulation. However, the primacy of seed paring can be explained by how the protein of the silencing complex (Argonaute) presents the 5' region of the miRNA preorganized to prefer pairing to the mRNA. To enhance both the affinity and specificity for matched mRNA regions, the RISC should present nucleotides 2-8 of the miRNA preorganized in the shape of an A-form helix to the mRNA[256,261].

MicroRNA binding sites that are conserved across species are much more likely to be biologically functional. The use of conserved binding sites reduces the false positive rate

of prediction tools significantly. However, there are many mRNAs with non-conserved 7 nt sites for each miRNA and the set of mRNAs that are coexpressed with a miRNA constitute a large number, yielding the possibility for much non-conserved targeting [262]. Thus, target prediction tools without any considering site conservation [260,263] or both with and without conservation cutoffs [257,264,265] have been developed.

Considering the thermodynamic stability by using the free energy of a miRNA-target duplex ($\triangle G_{duplex}$) is also important in the miRNA target prediction. An energetically more stable state is given when two complementary RNA strands are hybridized. The lower the free energy of two paired RNA strands (miRNA-mRNA), the more energy is needed to separate this duplex formation. Therefore, a miRNA has a higher affinity to bind to a mRNA, when the resulting RNA duplex has a low free energy. Moreover, for the identification of miRNA targets the secondary structure of mRNAs should be considered. The target site has to be accessible (open or unpaired) for miRNA binding, revealed by a defined energetic cost $\triangle G_{open}$. Further, additional nucleotides upstream and downstream of the target site, respectively, are also required to be unpaired [264]. The total free energy change, $\triangle\triangle G$, of the binding process is determined by the difference between the free energy gained by the miRNA-mRNA binding, $\triangle G_{duplex}$, and the free energy lost by unpairing the target-site nucleotides, $\triangle G_{open}$, and represents an energy-based score for the accessibility of the target site and the probability for a miRNA-target interaction [264,266].

Not only sequences of target sites can explain much of targeting specificity but also the UTR context [256,260]. Features of the UTR context have influence on the site efficacy. For example, the site has to be located within the 3'UTR at least 15 nt from the stop codon and away from the center of long UTRs, because in the center the site might be less accessible to the silencing complex. Moreover, high local AU content near a site increases its accessibility due to the weaker mRNA secondary structure. These assumptions are supported by the analysis of orthologous 3'UTRs and conserved 7-mers in general [260,267]. In addition, proximity to binding sites of coexpressed miRNAs boosts site efficacy, as two sites that are close together (within 40 nt, but no closer than 8 nt) tend to act cooperatively [260,268].

3.6.2 Correlation to Expression Profiles

When transfecting miRNAs into cells or by their overexpression it has been shown that a large number of mRNAs are downregulated, which indicates that these mRNAs are likely targets to the individual miRNAs[46]. Therefore, lowly expressed genes within a tissue in which a specific miRNA is highly expressed are potential targets to the miRNA. MicroRNAs can have hundreds of different targets in a cell and, at low expression levels, the miRNA may have minimal impact on any one of its target genes. However, genome-wide computational and transcriptome analyses showed that the expression of miRNAs is more positively than negatively correlated with that of their targets[40,41]. Moreover, Arvey *et al.* hypothesize that miRNAs that have a higher number of available targets will downregulate each individual target gene to a lesser extent than those with a lower number of targets[42].

3.6.3 Prediction Tools

In this study three different target prediction tools have been used including *miRanda*, *PicTar* and *TargetScan(S)*. In general, the predictions given by different tools are diverse and the amount of overlapping miRNA-target predictions is quite small. Reasons for largely non-overlapping predictions are for instance the level of stringent seed pairing, alignment artifacts, the use of slightly different UTR databases, the use of different miRNA sequences or intrinsic to the prediction algorithms themselves such as the treatment of the target nucleotide opposite to the first miRNA nucleotide[256].

The target predictions available from *microRNA.org* are based on an implementation of the miRanda algorithm[269,270]. For each miRNA, target genes are selected on the basis of three properties: sequence complementary, free energies of miRNA-mRNA duplexes and conservation of target sites in related genomes. First, miRanda analyzes the sequence complementary between a given mRNA and a set of miRNAs using a position-weighted local alignment algorithm. A weighted sum of scores for matches and mismatches of base pairs is computed, thereby the weights are position dependent. G-U wobble base pairs are allowed but scored less than perfect matching base pairs. Scores for base pairing at positions 2-8 have a greater weight and, in addition, base pairings in the 3' regions are also weighted higher in regard to e.g. 3' compensatory matches. Second, the free energy of the miRNA-mRNA duplex is estimated by using the Vienna RNA folding approach[271]. Finally, the conservation of target sites based

on PhastCons score[272] is considered to filter out less conserved predicted targets. In any additional step, the target sites predicted by miRanda are scored for likelihood of mRNA downregulation using mirSVR[273], a regression model that is trained on sequence and contextual features of the predicted miRNA-mRNA duplex[269].

Another popular algorithm used for the identification of miRNA targets is PicTar. It identifies potential targets for single miRNAs and moreover, PicTar ranks target genes by considering whether the mRNA is targeted by combinations of miRNAs[274]. In each cell type different miRNAs are coexpressed, which suggest a tissue-specific target gene regulation. Therefore, PicTar needs a set of miRNAs and a group of orthologous 3'UTRs from multiple species to determine common targets for the miRNAs. These miRNAs are then ranked by their likelihood. For single miRNAs perfect 7mer seed matches (either nucleotides 1-7 or 2-8) are required. The results are then filtered by checking the conservation of target sites and evaluating the free energy of the miRNA-mRNA duplex using RNAhybrid[275]. To each remaining target site a probability score is assigned corresponding to their likelihood of being functional[274]. The final probability scores are used in the sequence scoring algorithm, which computes a maximum-likelihood score for each species using a Hidden Markov Model (HMM). The final (combinded) score describes the likelihood of a gene being target to the given miRNA set[274,276].

The first version of the TargetScan prediction tool searched for seed pairing and ranked the resulting sites by evaluating thermodynamic stability. The results for multiple species are combined to get the predictions for conserved target sites[265]. A more simplified method called TargetScanS was later published[257], which searched for pairing to a 6-nt miRNA seed with an additional base pair at nucleotide 8 or a 1A-anchor. Furthermore, a method for evaluating site conservation was introduced[44] and target sites with imperfect seed matches but 3' compensatory pairing are also predicted. In mammals the efficiencies of the target sites are assessed by observing the UTR context of the target sites[260].

3.7 Analysis of Genomic Sequence Alterations

3.7.1 Identification of Local Variations

In the last years, several computational strategies have been developed to identify local (SNVs and InDels) and structural (e.g. CNVs) variations after mapping DNA sequencing reads to the reference genome. This study implemented the identification of local variations using two applications, namely the Roche GS Reference Mapper[57] and VarScan[277] for 454/Roche pyrosequencing reads and Illumina sequencing-by-synthesis reads, respectively.

The GS Reference Mapper (Newbler) application aligns pyrosequencing reads against a reference sequence and generates consensus sequences of the reads that align against the reference. In addition, Newbler also computes statistics for variations found in the reads, relative to the reference, and evaluate these lists of putative variations to identify so-called high-confidence nucleotide differences (HCDiffs). The application uses a combination of flow signal information, quality score information and difference type information to determine if a difference is high-confidence. In general, there must be at least 3 non-duplicate reads with at least one from the forward and reverse strand showing the difference, unless there are at least 5 reads with quality scores over 20 or 30 if the difference involves a homopolymer of 5 or more nucleotides[57]. Pyrosequencing uses the fluorescent signal strength of incorporated nucleotides in a homopolymer to estimate its length. The signal strength for homopolymer stretches is only linear for up to eight consecutive nucleotides, resulting in a higher error rate for larger homopolymer stretches[278]. However, the usage of the flow signal information in the Newbler application significantly improves resolving homopolymeric stretches of a sequence and thus, for pyrosequencing reads, Newbler performs better for SNV and InDel calling than all other methods.

Given a file with read alignments, the VarScan application scores and sorts the alignments on a per-read basis, discarding reads that aligned with low identity or to multiple locations in the reference sequence. The single best alignment for each read is then checked for sequence variations and variations detected in multiple reads are combined together into unique SNVs and InDels. For each predicted variation, VarScan determines the overall coverage, the number of supporting reads, average base quality and number of strands observed for each allele. After filtering, in which thresholds for coverage, quality, etc. can be set automatically or manually, VarScan reports SNVs and

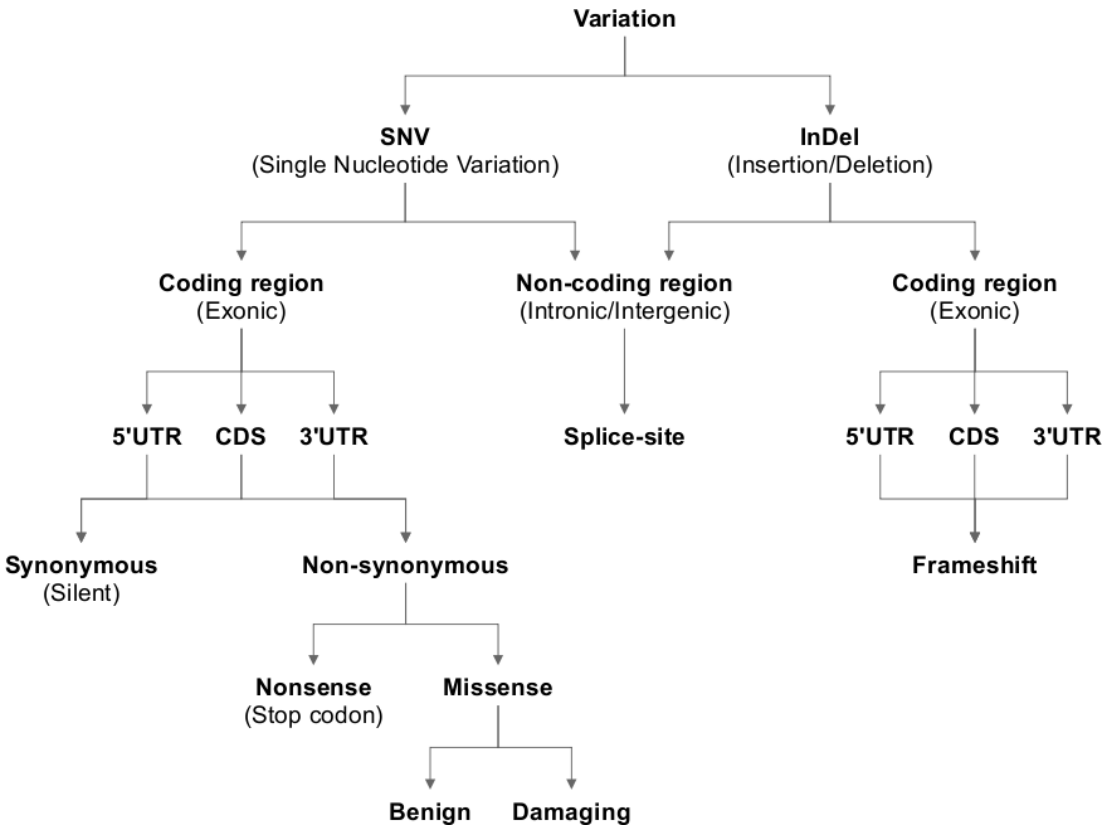

Figure 3.7: Annotation and functional characterization of local variations.

InDels with their chromosomal coordinates, alleles, flanking sequence and supporting read counts[277].

3.7.2 Annotation and Functional Characterization

After local variation calling one of the first postprocessing steps is their annotation and functional characterization. Variations are annotated based on different resources including databases from UCSC[219], NCBI[208] (e.g. Genbank[279], dbSNP[280] and OMIM[281]), ENSEMBL[207] and UniProt from EBI[282]. The annotation includes genomic locations (exonic, intronic, intergenic), gene names and their exonic locations (5'UTR, CDS, and 3'UTR important for miRNA binding), dbSNP entries (known or novel variations), SNV and InDel functions (nonsense, missense, frameshift, splice site affecting), protein positions and amino-acid changes, conservation scores (e.g. PhastCons[272]) and clinical associations (e.g. OMIM). For all missense SNVs it is possible to predict whether

they reside in an amino acid substitution affecting the protein function. For example, PolyPhen-2 is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using physical and comparative considerations[283]. Another prediction tool is SIFT, which is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences[284]. An overview about the genomic annotations and possible functional characterization of local variations is given in Figure 3.7.

3.7.3 Filtering

To find out which of the identified local variations might be functional and moreover, to reduce the search space of possible disease associated variations, different filtering steps can be applied. Most straightforward is the filtering by coverage, supporting reads, variation allele frequency, average base quality and supporting strands. For example, if the allele frequency range is 20-80% the variation is called heterozygous, and for more than 80% homozygous[51], meaning that local variations with less than 20% allele frequency might not be functional and thus, should be filtered out.

In general, false positives during local variation calling arise from two phenomena, sequencing errors and alignment artifacts. Errors on the Roche/454 platform are not dependent on read position, but tend to cluster around homopolymeric sequences that are often under- or overcalled[278], resulting in reads that contain gaps relative to the reference sequence. The second origin are alignment artifacts due to relatively short read length from NGS platforms and complexity of the (human) reference genome. For example, paralogous sequences and low-copy repeats that differ by only few bases can give rise to reads that, when aligned incorrectly, appear to support a local variation at the same position. These errors can manifest even in regions of high coverage. A window-based filtering approach that identifies clusters on SNV calls (i.e. three SNVs within 10 bp) might be useful to remove some of these artifacts[285]. Moreover, local variations with excessively high read depth are usually caused by structural variations or alignment artifacts and should also be filtered out by, for example, setting a maximum read depth according to the average coverage.

Finally, variation not predicted to be damaging, nonsense, frame-shifting or splice site affecting can also be removed because they might not be functional. In addition, the final set of filtered variations can be subsequently reduced to novel variations using

dbSNP annotations or annotations of polymorphic regions from other projects such as the *1000 Genomes Project*[286] or the Danish exome resequencing project[287]. Filtering for variations not in dbSNP can reduce the search space by 2-10 fold. However, when discarding known variations, in general rare variations, which might be pathogenic or known to be disease associated, are also filtered out. Therefore, variations with a known MAF of less than or equal to 0.01 or known disease associated variations present in the OMIM database were retained in this study.

Chapter 4

The Cardiac Transcription Network Modulated by the Transcription Factor Srf, Histone 3 Acetylation, and MicroRNAs

4.1 General Purpose and Previous Analysis

In this study we investigated the interplay of transcription factor binding, co-occurring histone modifications and miRNAs in regulating cardiac transcription networks. The aim was to understand how these molecular levels are involved in regulating cardiac transcription profiles and how they are connected to each other.

First, we focused on the four key TFs Gata4, Mef2a, Nkx2.5 and Srf and performed ChIP-chip experiments to determine their direct target genes. Several hundreds of TF binding sites could be identified for each factor (Chapter 2.2.1). Moreover, it has been shown that the four TFs analyzed have common binding pattern and can partially compensate each others function[1].

The expression of genes is mostly regulated by multiple TFs. To study the potential functional consequence of the frequent co-binding, siRNA knockdown experiments of the respective factors were performed in our group (data not shown). For Srf, we found 519 significantly differentially expressed transcripts in the siRNA-mediated knockdown, most of them being upregulated (in total 468) and only few downregulated (in total

51). Only transcripts that had a Benjamini-Yekutieli corrected p-value (see Chapter 3.5) of less than or equal to 0.05 in two knockdowns of Srf using different siRNAs when compared to non-specific siRNA (often called 'siNon') were considered to be significantly differentially expressed. The additional measurement of a non-specific siRNA is crucial for every siRNA experiment to access consequences on the cellular transcription profile that are caused by the RNAi experiment itself and not by the induced siRNA. In general all TFs are mainly transcriptional activators with 70-90% downregulated transcripts in siRNA knockdown. Most interestingly, genes bound by multiple factors are significantly less likely differentially expressed in siRNA knockdown than expected. This shows a buffering or compensation effect between the studied factors[1].

To investigate the influence of histone modifications as an epigenetic mechanism to modulate gene expression, we analyzed our TF binding data in correlation of co-occurring with four activating histone marks (H3ac, H4ac, H3K4me2/3; Chapter 2.2.1). With the focus on H3ac we found that ∼60% of observed histone 3 acetylation co-localize with binding events of the studied transcription factors. This is significantly more than what would be expected in a random situation (i.e. only 23% are expected to co-occur). Further, it was shown that the presence of H3ac marks has a significant impact on target gene expression. Genes marked by Mef2a or Nkx2.5 show significant increased expression levels compared to non-marked genes independent of co-occurrence of H3ac or not. In contrast, target genes directly bound by Gata4 or Srf were only significantly higher expressed when they were additionally marked by H3ac[1]. The Srf cofactor Myocardin has been reported to recruit histone acetyltransferase p300 to Srf binding sites whereby H3ac is induced and gene expression enhanced[90].

Our previous analyses are based on ChIP-chip as well as siRNA knockdown experiments (see Schlesinger *et al.*[1] for more information). To validate and further investigate the correlation of H3ac and Srf target gene expression, we performed genome-wide ChIP-seq experiments in HL-1 cardiomyocytes in our group. The resulting ChIP-seq data (Chapter 2.2.1) have been analyzed in this study and the results are shown in the following. Most interesting, it is shown that H3ac tags have the potential to buffer downregulation of direct Srf targets in an siRNA mediated knockdown. In addition, the influence of miRNAs on the Srf driven regulatory network is shown.

	Srf		H3ac	
Total number of sequenced reads	6,967,318		8,364,328	
Number of low quality reads	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 reads	2,266,839	(33%)	2,039,627	(24%)
Number of called peaks	2,190		10,486	

Table 4.1: Number of ChIP-seq read matches and called peaks for Srf and H3ac. Only uniquely mapped reads without any error (perfect matches), with one error (1-error matches) and with two errors (2-error matches) were retained for peak calling. Reads that could not be mapped to the mouse genome or mapped to multiple genomic locations (summarized as unmatched reads) or with of low quality (containing one or more ambiguous bases) were discarded from further analysis. Percentages are computed in respect to the total number of sequenced reads.

4.2 Analysis of ChIP-seq Data for Srf and Histone 3 Acetylation

To confirm and further investigate the impact of H3ac on Srf target gene expression, ChIP-seq experiments were performed using HL-1 cardiomyocytes measuring Srf binding and histone 3 acetylated sites on a genome-wide scale. Deep sequencing of the individual ChIP experiments resulted in 6,967,318 and 8,364,328 reads for Srf and H3ac, respectively (see Chapter 2.2.1). Thereof, 4,543,634 reads (65.2%) for Srf and 6,141,144 reads (73.4%) for H3ac could be mapped to the mouse reference genome (NCBI v37; mm9) using the read mapping tool RazerS[174] (Chapter 3.1). Only uniquely mapped 36 bp reads with at most two mismatches were retained for peak calling. The mapping results indicate good experimental qualities (error distribution of reads for both experiments is given in Table 4.1).

To identify Srf and H3ac binding sites, the one-sample approach implemented by the CisGenome[134] software (Chapter 3.2.1) was used for several reasons. Most importantly, no Input sample was measured and thus, the used peak calling algorithm had to estimate the null distribution from the ChIP sample itself. For the Srf ChIP-seq data CisGenome 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, ensuring a FDR lower than 2% for significant called peaks. As histone enriched sites were shown to be broader than transcription factor peaks, a window size of 250 bp, a step size of 50 bp for the sliding and a minimal

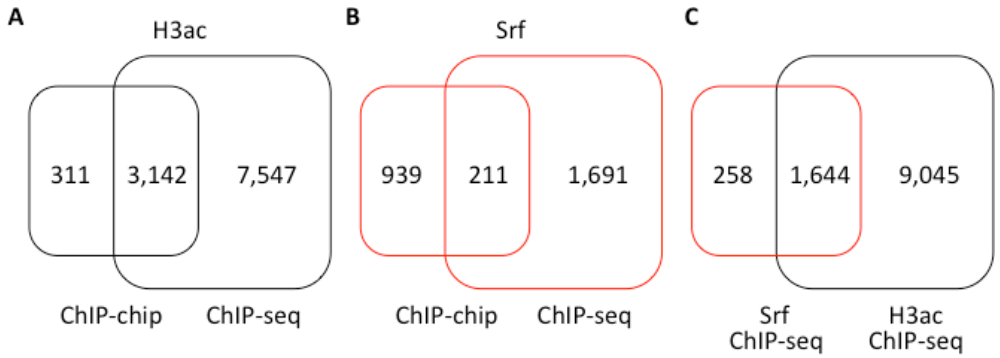

Figure 4.1: Identified target genes of Srf and H3ac in ChIP-chip and ChIP-seq. (A+B) Overlap between genes associated to (A) H3ac and (B) Srf peaks in ChIP-chip compared to ChIP-seq. (C) Overlap between Srf and H3ac target genes in ChIP-seq.

read count level of 10 was used for the H3ac ChIP-seq data, ensuring a FDR lower than 5% for significant called peaks.

After the peak calling procedure, the described boundary refinement (Srf and H3ac) and single-strand filtering (only H3ac) were applied. After manual inspection of individual peaks, application of single-strand filtering for the Srf ChIP-seq data was omitted as it resulted in a great loss of Srf binding sites because the majority of peaks was not equally represented on the 3' strand. The most likely reason for this is an insufficient shearing of the genomic DNA during the ChIP procedure leading to DNA fragments of non-optimal size (typically in range of ∼150-300 bp[134]) for the sequencing. To substantiate this assumption the resulting fragments were further analyzed by gel electrophoreses. The gel showed a heterogenous size distribution with a proportion of fragments longer than optimal for ChIP analysis (data not shown). Finally, the ChIP-seq approach identified 2,190 and 10,486 peaks for Srf and H3ac, respectively, on the whole mouse genome (Table 4.1).

4.2.1 Comparison of ChIP-seq versus ChIP-chip

As the ChIP-seq and the ChIP-chip approach both aim to measure the same enriched binding sites but use different techniques with different sensitivities, the overlap based on target genes between these two techniques was analyzed.

For ChIP-seq, 1,902 and 10,689 target genes were defined to be associated to the identified peaks (described above) for Srf and H3ac, respectively. In contrast to the genome-

wide ChIP-seq approach, as expected, a much lower number of target genes and associated peaks were found in ChIP-chip (see Chapter 2.2.1). In total, 1,150 and 3,453 target genes were associated to Srf and H3ac peaks, respectively, in ChIP-chip. Out of the 3,453 target genes associated to ChIP-chip H3ac peaks, 91% overlapped with the ChIP-seq data (Figure 4.1A). However, for the 1,150 genes associated to ChIP-chip Srf peaks the overlap was only 18% (Figure 4.1B). Finally, most (86%) of the Srf target genes were found to have an additional H3ac modified site (Figure 4.1C).

4.2.2 Confirmation of Histone 3 Acetylation Dependent Expression of Srf Targets

Based on ChIP-chip, it was shown that the presence of H3ac marks has a significant impact on Gata4 and Srf target gene expression (see Chapter 4.1). To validate and further investigate the correlation of H3ac with Srf target gene expression, we analyzed the ChIP-seq data in the same way as the ChIP-chip data, despite the differences in the actual peaks (described above). In summary, we found a similar synergistic effect of H3ac and Srf binding when compared to non-bound genes or genes solely bound by either of both (Figure 4.2A).

The influence of H3ac marks was further substantiated by integrating the ChIP-seq results with the RNAi knockdown data (described in Chapter 4.1) of Srf in HL-1 cells. In accordance to its mainly activating function, we found a significant decrease in expression levels of genes bound by Srf without any H3ac marks. However, this decrease was significantly smaller in genes that were additionally marked by H3ac in the wildtype pointing to a buffering effect of H3ac on Srf target gene expression after reduction of Srf protein (Figure 4.2B).

4.3 Impact of MicroRNAs on the Srf-Driven Transcription Network

Considering that only a small proportion of differentially expressed genes in loss-of-function experiments are direct targets of the respective transcription factors, we studied the potential impact of miRNAs. We asked whether the transcription factor Srf regulates miRNAs, because Srf is known to regulate cardiac-relevant miRNAs like miR-1 and miR-133[95,169].

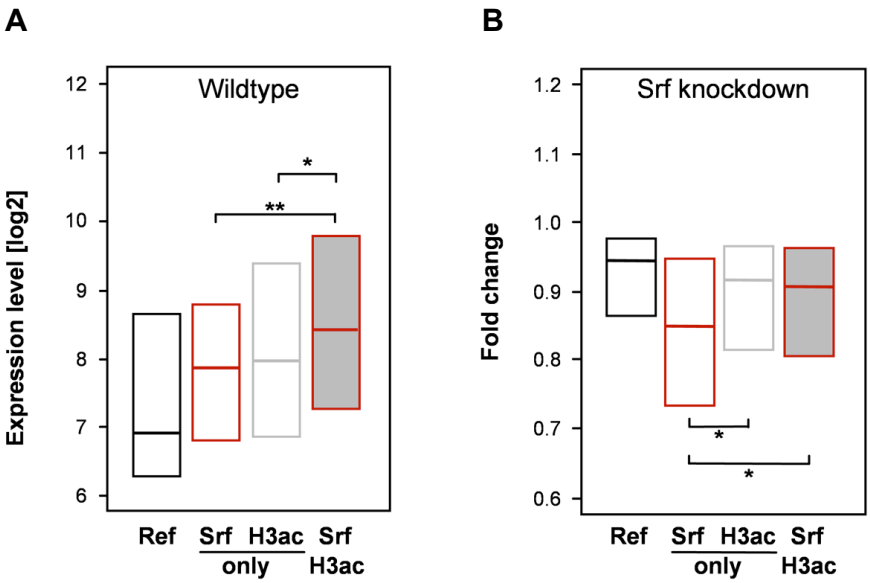

Figure 4.2: Confirmation of H3ac dependent expression of Srf targets by ChIP-seq. (A) Boxplots of expression levels of transcripts grouped according to H3ac and/or Srf binding close to the transcriptional start site (TSS $<$ 1.5 kb). (B) Boxplots of fold changes relative to siNon control (non-specific siRNA) of downregulated transcripts after Srf knockdown grouped according to H3ac and/or Srf binding close to the transcriptional start site (TSS $<$ 1.5 kb). (A+B) Genes showing neither binding of investigated transcription factors nor H3ac are used as reference. The resulting p-values are indicated: $p<0.01$ (**) and $p<0.05$ (*).

For analyzing the direct regulation of miRNAs it was not possible to use or integrate the Srf ChIP-chip data, as it relies on a pre-designed array, which was built to represent gene but not miRNA promoters. Therefore, we used the Srf ChIP-seq data (Chapter 2.2.1) to detect direct Srf regulation of miRNAs. Moreover, for analyzing the indirect regulation of miRNAs, we used the miRNA-seq data described in Chapter 2.2.2.

First, to find direct Srf regulation of miRNAs, the 2,190 peaks from the ChIP-seq experiment were used to map binding sites of Srf potentially regulating miRNAs. Using the known miRNAs annotations in mouse (retrieved from the miRBase[43] database v14.0) 22 miRNAs were predicted with a direct Srf binding site within a region of 10 kb (based on the ChIP-seq peaks). Among these miRNAs, we found several well-known cardiac-relevant miRNAs like miR-1, miR-125b, miR-133, miR-143 and miR-145 (Supplementary Table S1).

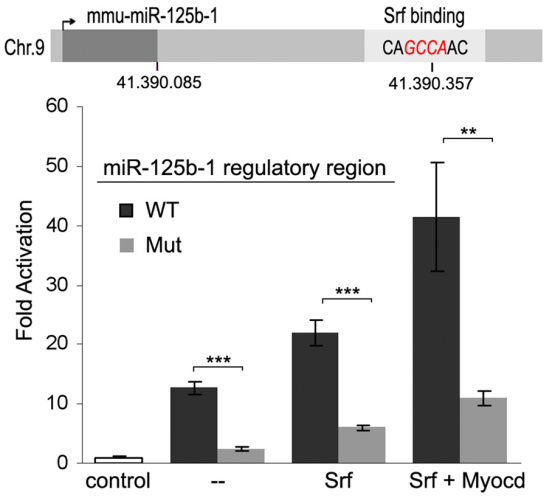

Figure 4.3: Promoter Analysis of mmu-miR-125b-1. Srf ChIP-seq analysis revealed an Srf binding region downstream of mmu-miR-125b-1. Shown are the positions of mmu-miR-125b-1 and the Srf binding motif with its core sequence in red. The Srf ChIP-seq peak region was cloned as mmu-miR-125b-1 promoter into the pGL3basic vector for luciferase reporter gene assay. Srf alone and in combination with its cofactor Myocardin (Myocd) significantly increased the activation of the luciferase beyond activation driven by endogenous Srf. Mutation of the core sequence (GCCA to TAGT) of the Srf binding motif (Mut) abolished activation by Srf and Myocd compared to the wildtype (WT).

Out of the 22 miRNAs with direct Srf binding sites, one site in the regulatory region of mouse miR-125b-1 was selected and also experimentally validated in our group using luciferase reporter gene assays. Mmu-miR-125b is known to be deregulated in heart diseases[288] and was found to be differentially expressed in Srf siRNA knockdown. Figure 4.3 shows the Srf binding motif and respective Srf ChIP-seq peak within the regulatory region of miR-125b-1. Luciferase reporter gene assays with wildtype and mutated fusion constructs confirmed its functionality. Mutation of the potential Srf binding sequence (CAGCCAAC to CATAGTAC) significantly reduced the transcriptional activity of the reporter gene.

Second, to study if a significant reduction of the Srf protein in cardiomyocytes would affect the expression of associated miRNAs (indirect regulation), another siRNA experiment was carried out again using two siRNAs against Srf (Srf siRNA-1/2) and one non-specific siRNA (siNon) but now followed by miRNA quantification using next-generation sequencing (Chapter 2.2.2). In this study, MicroRazerS (Chapter 3.1) was

used to map the sequenced reads to the mouse reference genome (NCBI v37; mm9). An evaluation of MicroRazerS in comparison with other short read mapping tools based on small RNA reads from human heart samples (Chapter 2.2.3) is given in Chapter 4.3.1. Using MicroRazerS with a seed length of 16 bp, a maximal number of 20 best hits and at most one mismatch and no InDels in the seed length resulted in 5,449,988 (96.7% for Srf siRNA-1), 5,296,564 (96.2% for Srf siRNA-2) and 5,475,045 (96.5% for siNon) unique read sequences that could be mapped to the mouse genome representing 97.3% (14,504,934 for Srf siRNA-1), 96.8% (14,053,178 for Srf siRNA-2) and 97.1% (14,307,881 for siNon) of the sequenced reads. The seed length of 16 bp was found to be optimal when searching for miRNAs which have a length of 19-25 nucleotides. Using the annotation from the miRBase[43] database (v14.0), the reads could be mapped to 349 (Srf siRNA-1), 365 (Srf siRNA-2) and 363 (siNon) known miRNAs. Using the miRNA-seq approach followed by the described mapping process, in total 370 miRNAs could be identified. To subsequently test if any miRNA showed differential expression between siNon and Srf knockdown, Fisher's exact test was applied comparing the number of reads mapped to a single miRNA between the siNon and the knockdown samples normalized by the total number of reads that could be mapped to any miRNA in the respective samples. Using a Benjamini-Yekutieli corrected p-value (see Chapter 3.5) of less than or equal to 0.05 as significance threshold, 42 miRNAs (49 loci) were found to be differentially expressed in both siRNA knockdown experiments, including heart-relevant miRNAs such as miR-208, miR-125b and miR-21 (Supplementary Table S2). We found that most of the significantly differentially expressed miRNAs (78%) were downregulated supporting the role of Srf as a miRNA activator.

To explore the potential regulatory effect of differentially expressed miRNAs on the Srf network, miRNA target prediction was performed for 77 differentially expressed miRNAs including the 42 previous miRNAs and 35 additional miRNAs that were differentially expressed in only one sample (Srf siRNA-1 or Srf siRNA-2, respectively). The miRanda[269,270] algorithm (Chapter 3.6.3) was applied to 3'UTR sequences. Very restrictive parameters including a score cut-off ≥ 140 (default = 50), a gap open penalty of -9 and a gap extension penalty of -4 were used to ensure a low number of false positives. Using these parameters, the target prediction revealed 192 of 429 differentially expressed genes to be potential direct targets of a differentially expressed miRNA. Applying Fisher's exact test, this number was found to be significant when compared to all possible target genes ($p = 1.77 \times 10^{-5}$). Compared to all predicted differentially expressed genes, we found a higher fraction of upregulated genes (57% of all upregu-

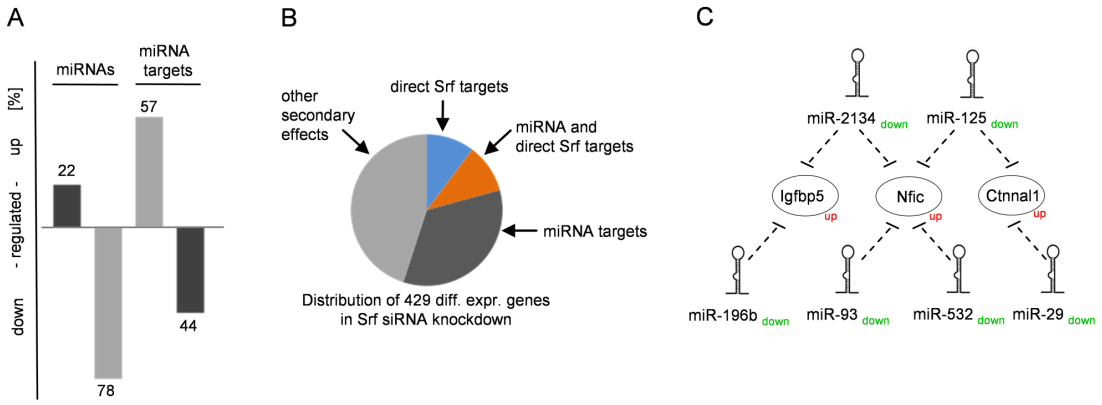

Figure 4.4: Impact of miRNAs on the Srf-driven cardiac transcription network. (A) siRNA knockdown of Srf in HL-1 cardiomyocytes results in 77 differentially expressed miRNAs. Target gene prediction of these mostly downregulated miRNAs revealed 192 differentially expressed genes, with a higher fraction of upregulated genes (57% of all upregulated genes) than downregulated genes (44% of all downregulated genes). (B) Direct Srf targets represent only a small fraction of all differentially expressed genes in Srf knockdown (orange and blue). Targets of differentially expressed miRNAs impact 45% (dark gray) with a partial overlap of direct Srf targets (orange). Approximately 50% of differential expression is driven by other secondary effectors (light gray). (C) Exemplary network of potential indirect gene regulation by miRNAs. The genes Igfbp5, Nfic and Ctnnal1, which are no direct targets of Srf, are predicted targets for a set of downregulated miRNAs and are found to be upregulated in the Srf knockdown.

lated genes) compared to downregulated genes (44% of all downregulated genes) to be miRNA targets (Figure 4.4A).

Thus, the differential expression of miRNAs in the Srf knockdown has the potential to impact up to 45% of all differentially expressed genes directly. Nevertheless, the current miRNA target prediction tools are still quite unreliable to predict real regulatory dependencies with high accuracy. However, given that the miRNA targets found in this study might themselves be transcriptional regulators, miRNAs very likely provide a substantial explanation for the observed consequences on the transcriptional portrait (Figure 4.4B). A representative example of an indirect TF regulation through miRNAs is shown in Figure 4.4C. It comprises the three genes Igfbp5 (insulin-like growth factor binding protein 5), Nfic (nuclear factor I/C) and Ctnnal1 (catenin alpha-like 1). None of these genes has an associated direct Srf binding site in ChIP-chip/seq but all are found to be upregulated in the Srf siRNA-mediated knockdown experiment. Strikingly, all are predicted targets of several miRNAs downregulated in the same knockdown.

4.3.1 Evaluation of MicroRazerS

MicroRazerS has been developed within this study and to evaluate our short read mapping tool we used a dataset derived from three human normal heart samples (see Chapter 2.2.3). Deep sequencing of the small RNA library produced 9,286,222 sequenced single-end reads of 36 bases in length, yielding 2,402,361 unique (i.e. non-redundant) read sequences. These unique reads were mapped to the human genome (NCBI v36.1; hg18) using Mega BLAST[192], SOAP2[181], Bowtie[176] and MicroRazerS. The mapping results are shown in Table 4.2. The running time was measured on an AMD Opteron 2384 with 32 GB memory running a 64-bit Linux system. In the test setting, MicroRazerS was nine times (170 min) faster than Mega BLAST and 20 min slower than SOAP2 or Bowtie. However, SOAP2 took 84 min and Bowtie 206 min to build a BWT index for the human reference genome. Moreover, Mega BLAST and SOAP2 produced huge output files that need to be filtered, i.e. in both cases additionally ∼30 minutes were needed for post-processing and filtering after mapping.

	MicroRazerS	Mega BLAST	SOAP2	Bowtie
Running time (min)	24	194	6	5
Buidling index (min)	-	-	84	206
Output size (GB)	0.1	8.6	6.8	0.7
Memory usage (GB)	3.4	1.4	8.3	2.3
Unique sequence aligned	1,319,218	891,215	1,318,504	1,184,590
Mappable reads	7,743,516	7,001,832	7,742,266	7,410,239
Reads annotated as miRNA	5,819,189	5,746,588	5,819,184	5,667,027
Total number of miRNAs	381	372	381	372
- miRNAs with read count >150	101	96	101	99

Table 4.2: A query dataset of ∼2.4M non-redundant read sequences of length 36 bp representing a total of ∼9.3M reads was used. Using MicroRazerS the parameters were set as follows: *-m* 20 (maximum number of best matches), *-pa* (purge ambiguous reads having more than 20 equally best hits) and *-sL* 16 (seed length). A seed length of 16 bp (100% identity) was used for all mapping tools. In the case of MicroRazerS, no mismatches in the read prefix were allowed. For SOAP2, 20 mismatches in one read were allowed but only exact matches in the seed region. For Bowtie, a quality cutoff *-e 500* was used, which corresponds to allowing 20 mismatches, as each base quality in all reads was set to Phred score quality of 25. The resulting alignments except those from MicroRazerS were filtered to get the best (longest) hits with at most 20 positions in the human genome.

To annotate the sequence reads with known miRNAs, we checked for overlaps with positions of precursor hairpin miRNAs annotated by the miRBase database (release 13.0). Of note, MicroRazerS was able to map a higher number of reads than all other programs. While in this dataset almost no differences in miRNA predictions between SOAP2 and MicroRazerS were observed, the slightly lower sensitivity of SOAP2 could lead to missing miRNA measurement in other datasets. Allowing to map reads with at most one error in the seed sequence to be robust in the presence of possible sequencing errors and SNVs, we observe that indeed a higher number of reads can be annotated as miRNAs. Using this option, MicroRazerS mapped 97% of all unique sequences to the human genome representing 99% of the total reads, resulting in 414 known miRNAs.

4.4 An Srf Centered Transcription Network

In addition to a genome-wide perspective, the analysis also provides useful information on the level of individual genes. An extensive literature search was conducted and an Srf centered transcription network was build, where the findings from the Srf and H3ac ChIP-chip/seq and Srf siRNA-mediated knockdown experiments were subsequently integrated (Figure 4.5). Thus, our data add regulatory content to the nodes, which are connected by referenced interactions. The network shows the common regulation by Srf and H3ac as well as the impact of the post-transcriptional modulation of expression levels by miRNAs. Target genes important in the cardiovascular context are grouped according to their biological roles like 'regulation in muscle contractility' or 'cardiac growth' and 'cardiac conduction'. As an example for the interplay between these different regulatory levels, the apoptotic machinery is regulated at all three levels (direct Srf binding, H3ac and miRNA post-transcriptional modulation) through several pathways involving pro-apoptotic (Casp3, miR-320, Hsp20/a8/a5, Bax) as well as anti-apoptotic (miR-21, Bcl2, Mcl1) regulators.

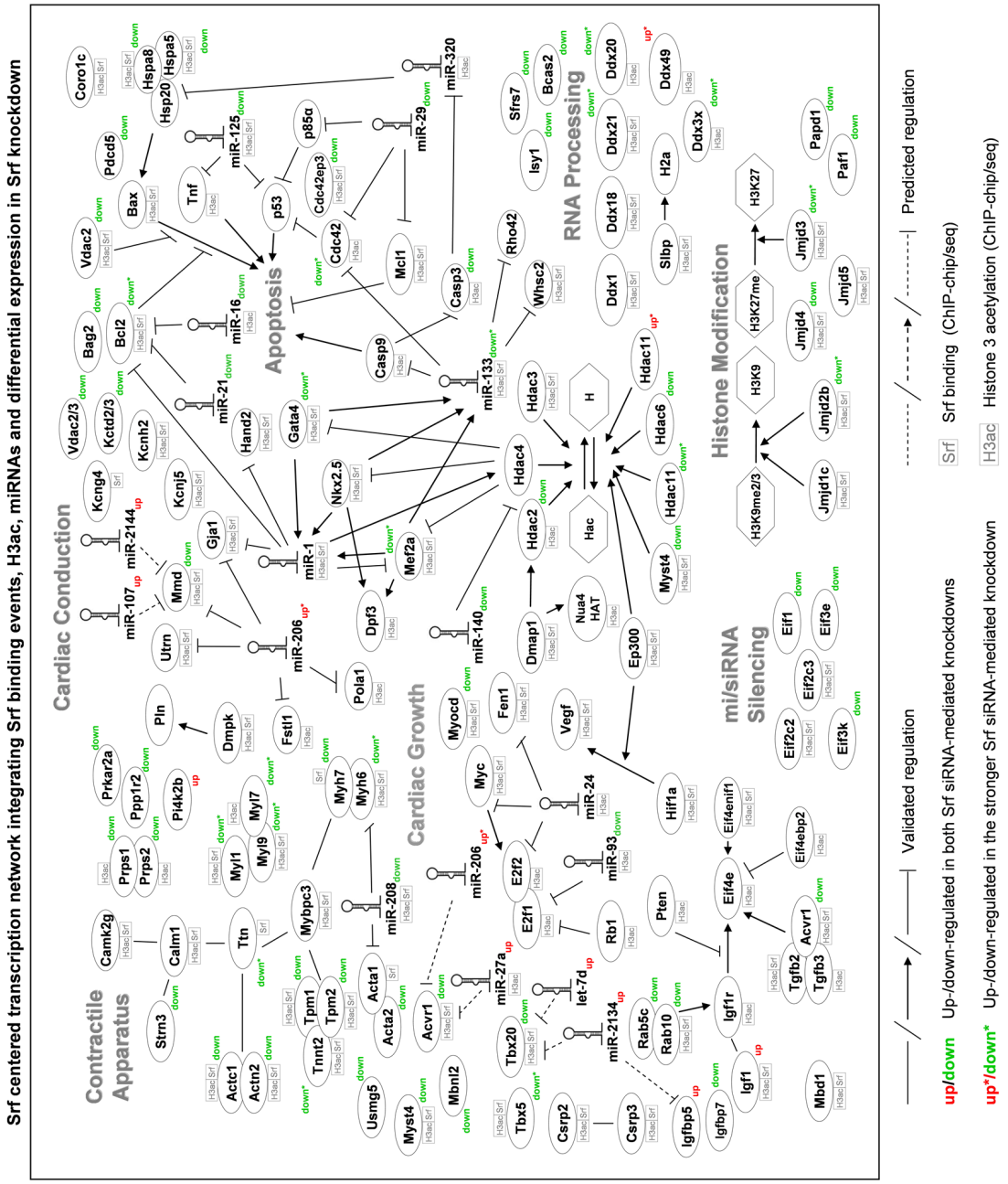

Figure 4.5: Srf centered transcription network integrating Srf binding events, H3ac, miRNAs and differential expression in Srf knockdown. The shown transcription network is based on an extensive literature search[1] and integration of our own findings. Data based on Illumina expression arrays, ChIP-chip/seq, miRNA-seq and qPCR. Srf binding and H3ac occurrence are depicted in small boxes and up (red) or down regulation (green) in Srf knockdown is further indicated.

Chapter 5

Dissecting Congenital Heart Disease - Genomic Sequence Alterations, Gene Expression and MicroRNA Profiling in Patients with Tetralogy of Fallot

5.1 General Purpose

Tetralogy of Fallot (TOF) accounts for 7-10% of all congenital heart disease (CHD), which are the most common birth defects in human. Considering the background hypothesis of CHD, most of them are likely caused by a panel of genetic variations with each effecting protein function or expression only modestly and manifest as disease only when combined with additional genetic, epigenetic or environmental alterations.

In the past, the discovery of oligogenic disorders has been less amenable to conventional genetic techniques. In this study we used next-generation sequencing techniques to discover sequence alterations in over thousand heart- and muscle-relevant genes and miRNAs in patients with TOF, parents and controls. The genetic architecture of TOF with an oligogenic mutation pattern is shown characterized by a combination of inherited and novel, common and rare alleles showing a high dependency of functionally interacting yet individual mutations. Further, we investigated genome-wide mRNA and miRNA levels in TOF patients and healthy unaffected individuals and combined gene

expression profiles with miRNA target predictions.

5.2 The Genetic Basis of Tetralogy of Fallot

To identify genomic sequence alterations and to analyze a potential oligogenic basis of TOF, we performed targeted resequencing of 18 patients with TOF of which 13 are unrelated sporadic cases and five are members of distinct families with recurrent CHD (see Chapter 2.2.4, Figure 2.5). To study the pattern of inherited and novel mutations we additionally sequenced nine family members consisting of seven healthy parents and two siblings affected with dextro-transposition of the great arteries (d-TGA) and tricuspid insufficiency (TI). The samples were sequenced by the 454 GS FLX instrument from Roche/454 and the Illumina GAIIx. On average sequencing resulted in ∼13,271,000 read pairs and ∼759,000 single-end reads per sample for Illumina and Roche/454, respectively (see Chapter 2.2.4, Table 2.2). Reads resulting from Illumina sequencing were mapped to the human reference genome (NCBI v36.1; hg18) using the Burrows-Wheeler alignment (BWA) tool[177] v0.5.9 with '*sampe*' command and default parameters. SNV and InDel calling was performed using VarScan[277] v2.2.3 with a minimum of three supporting reads, a minimum base quality of 20 (Phred score) and a minimum variant allele frequency threshold of 0.2. Mapping as well as SNV and InDel calling for reads resulting from Roche/454 sequencing were performed using the Roche GS Reference Mapper (Newbler) v2.5.3 with default parameters leading to high confidence differences (HCDiffs). Ambiguously mapped reads were discarded from the analysis either using Samtools[289] v0.1.12a in case of Illumina reads or the Newbler software for 454 reads. On average ∼9,744,000 (73.4%) read pairs (36 bp) and ∼755,000 (99.5%) single-end reads (∼400 bp) per sample for Illumina and Roche/454, respectively, were mapped to the human reference genome, with high average base quality and read coverage (Supplementary Figure S1).

Additional filtering of found local variations was performed for both techniques to ensure a minimum variant allele frequency threshold of 0.2 and a minimum coverage of five and ten sequenced reads for Roche/454 and Illumina, respectively. Moreover, sequence variations were functionally annotated using SIFT[284] and PolyPhen-2[283]. Afterwards, we filtered for local variations predicted to be missense, nonsense, frame-shifting, or affecting splice or miRNA binding sites. Only those missense SNVs were retained, which were predicted to be damaging or unknown, while tolerated variations were discarded. The final set of filtered variations was subsequently reduced to novel variations

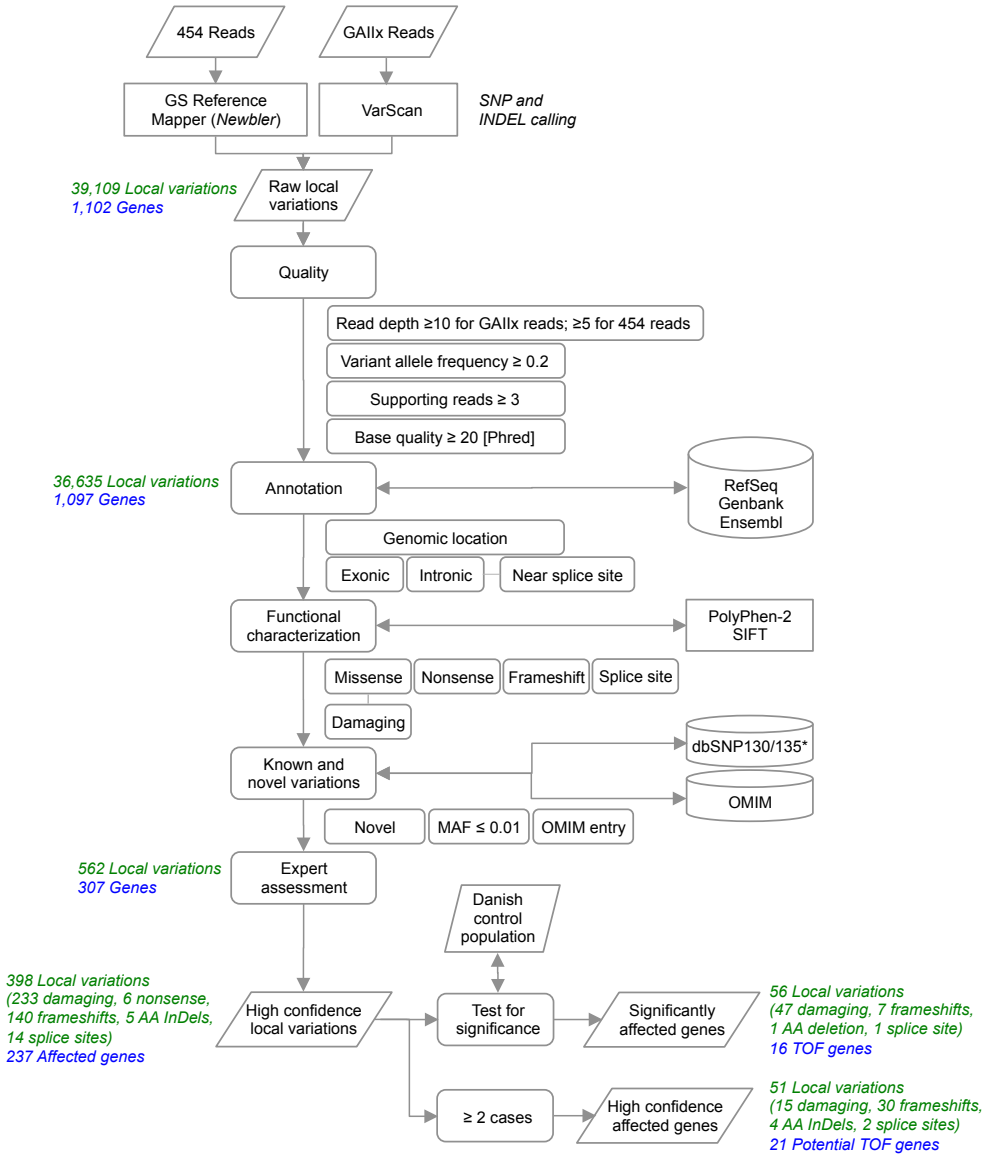

Figure 5.1: Filtering pipeline for local variations. 454 and GAII reads were mapped and used for SNP and InDel calling. After quality control, variations were functionally annotated, filtered and reduced to novel variations, variations with a minor allele frequency of less than or equal to 0.02 using dbSNP (v130) and known disease-associated variations. After manual assessment, high confidence local variations were statistically tested against the control population. *Variations in TOF genes were additionally examined using dbSNP (v135).

or variations with a MAF of less than or equal to 0.01 using dbSNP[280] (v130) annotations. Known disease associated variations present in the OMIM[281] database were retained irrespective of their MAF. To ensure high confidence, the local variations of TOF patients and family members were further manually assessed for potential biological function comprising gene as well as protein annotations, splice site alterations, technical biases and effective amino acid changes. The individual filtering steps are shown in Figure 5.1.

As a quality control we compared the gene length to the number of called SNVs and found no obvious correlation, meaning that some short genes have a high number of unique SNVs while long genes can have only few SNVs (Supplementary Figure S2). Furthermore, affected genes are equally distributed over all chromosomes (Figure 5.2). To technically confirm the genomic variations and to gain insights into the respective gene expression profiles in the heart, we gathered mRNA profiles from right ventricles of 22 patients with TOF as well as four healthy individuals. The description of the mRNA datasets and the gene expression analysis is given in detail in Chapter 2.2.4 and Chapter 5.3, respectively. We gathered all mRNA-seq reads which mapped to found local variations. A variation was defined to be validated if at least one mRNA-seq read mapping to the same genomic location showed the identical sequence alteration. Thereby, we were able to validate approximately 76% of local variations covered by mRNA-seq (on average over all individuals). Increasing the minimal mRNA-seq coverage resulted in an increased number of validated local variations. For example, using a minimal coverage of ten mRNA-seq reads ∼96% of local variations could be validated (Supplementary Figure S3 and S4).

Copy number variations (CNVs) have been examined within the ten TOF samples pyrosequenced by the Roche/454 technology. CNV calling for the long 454 reads resulting from Roche/454 sequencing was performed using the Roche GS Reference Mapper application v2.5.3 with default parameters resulting in high confidence rearrangement points and regions. High confidence structural variations (HCStructVars) were further filtered to be novel and manually assessed for biological function.

To enable the statistical assessment of found sequence variations, a Danish exome SNV dataset was incorporated as a large control cohort comprising 200 individuals (controls) of close genetic origins to the analyzed German individuals[287]. Using such a close control population is mandatory as exonic variations below 1% allele frequency show a high population-specificity[286]. Moreover, the selected control dataset further shows

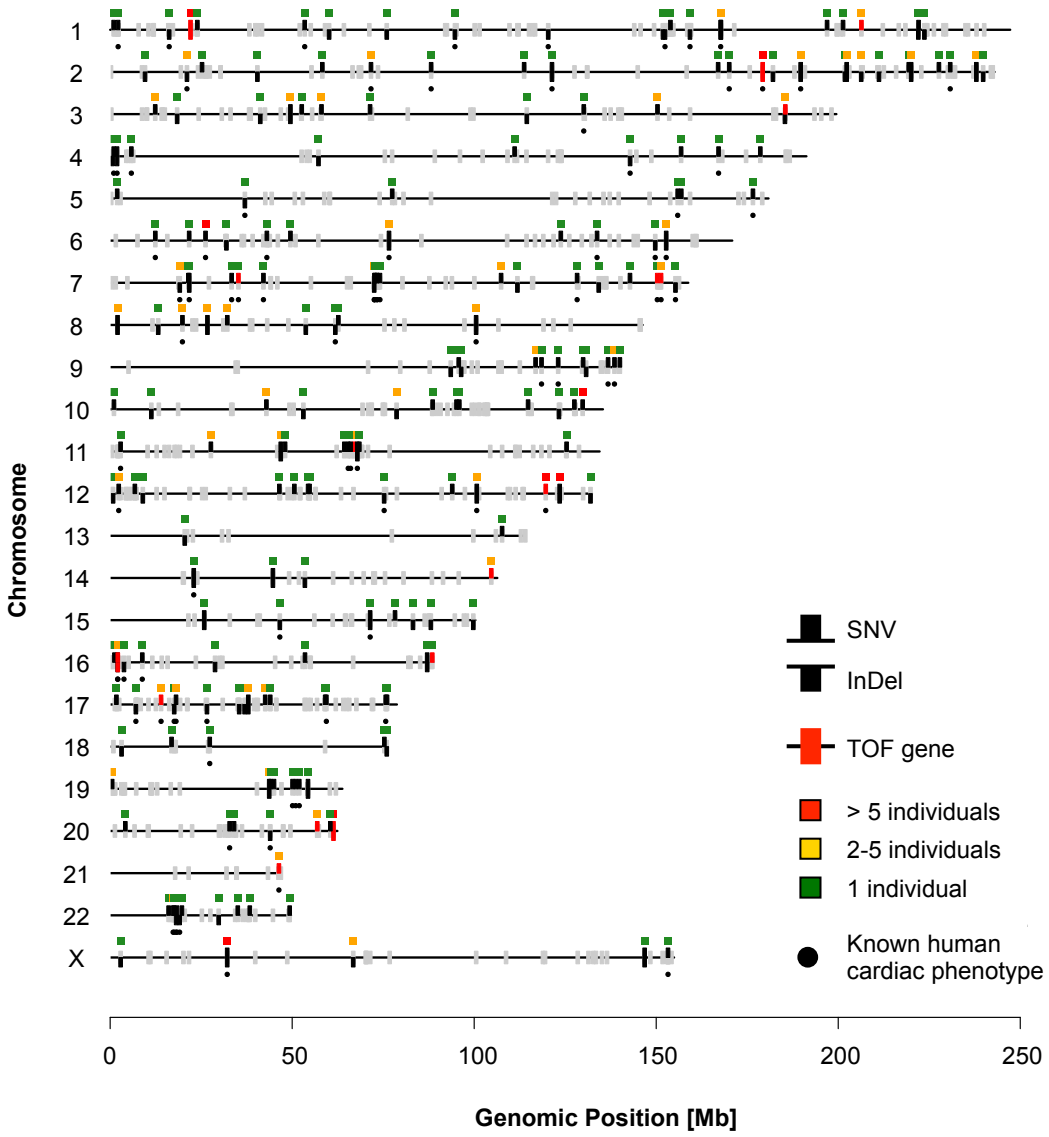

Figure 5.2: Genomic positions of affected genes. Chromosome length is represented by horizontal lines. Heart- and muscle-relevant genes initially selected for the study are shown in gray. The final set of genes with detected SNVs and InDels are marked by a black bar above or below the line, respectively. The 16 defined TOF genes are shown in red. The box above each affected gene indicates the number of TOF patients, which have at least one local variation in that gene. Dots below genes indicate known human cardiac phenotypes curated from literature.

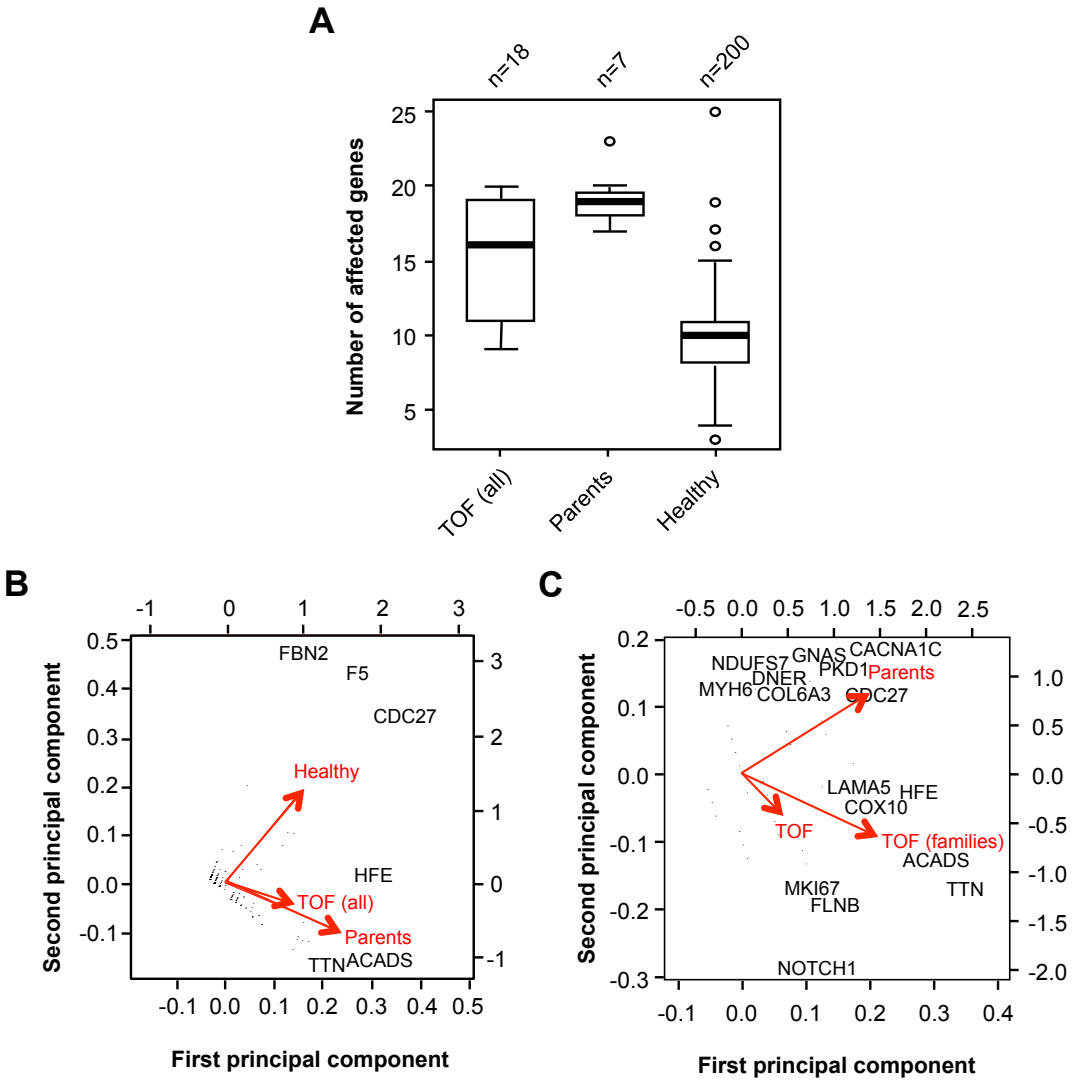

Figure 5.3: (A) Boxplot based on the number of affected genes in patients with Tetralogy of Fallot [TOF (all)], their parents and control individuals [healthy]. Data is based on SNVs only. TOF patients and their parents are enriched for genes with SNVs compared to healthy individuals. (B-C) Biplot of principal component analysis based on gene-wise SNV frequencies for patients with Tetralogy of Fallot (TOF), the analyzed parents (Parents) and the control population (Healthy). Genes with a high distance from zero in both components are indicated by their name. (B) Principal component analysis based on all three groups. (C) Principal component analysis based on TOF patients and analyzed parents. The patients have further been divided into individuals taken from the analyzed families [TOF (families)] and all other TOF patients [TOF].

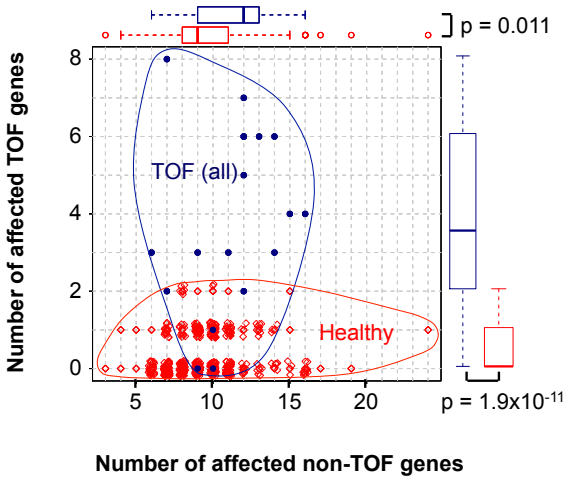

Figure 5.4: Scatterplot showing the number of genes with SNVs in respect to the defined set of significantly affected genes (TOF genes) versus all other genes (non-TOF genes). TOF patients and healthy individuals are represented by blue dots and red diamonds, respectively. TOF patients show a clear enrichment in mutations in the selected gene set. Distributions over the number of affected TOF and non-TOF genes per individual are given as box-plots. P-values are based on Wilcoxon rank sum test.

high similarity in experimental and analytical procedure ensuring high comparability. The retained total number of SNVs in this control population was subsequently filtered using the same pipeline established for our own variations.

SNV and InDel calling and filtering in TOF patients resulted in a total of 398 local variations altering the coding sequence of 237 genes classified as damaging (233), nonsense (6), frameshift (140) or splice site (14) mutations as well as amino acid InDels (5). CNV calling and filtering in ten TOF patients sequenced with Roche/454 technology resulted in three high confidence CNVs altering the coding sequence of three genes (Supplementary Table S3). No relevant mutations were observed in miRNA mature sequences, i.e. we found only few miRNA mutations and these are not located within the seed region (Supplementary Table S4). Variations in three genes (SGCA, MTPN and ZFPM2) were found in non-coding sequences related to predicted binding sites of five co-expressed miRNAs (hsa-miR-548j, hsa-miR-15a, hsa-miR-16, hsa-miR-195 and hsa-miR-873). These genes also showed genotype-specific expression in related cardiac biopsies (details are given in Chapter 5.5).

Further, the impact of differential splicing as a potential disease-causing mechanism

was evaluated. We found 1,765 significantly differentially expressed transcripts in TOF compared to normal heart (see Chapter 5.3), of which only 50 are related to differential splicing events. These transcripts were found with a different abundance (based on POEM estimation; see Chapter 3.3.1.1) between TOF and normal heart (RV) samples, i.e. with an average fold change greater than or equal to 2.0 or less than or equal to -2.0. Moreover, no deleterious sequence variations was found in a splicing factors. Looking at non-exonic mutations we found only few effective splice site mutations (Supplementary Table S5). Thus differential splicing is unlikely to be a TOF-associated mechanism.

In total, we found 237 deleterious mutations in genes of TOF subjects (based on local variations). 134 genes harbor exclusively SNVs, 36 genes SNVs and InDels and 67 genes only InDels. On average 16 and 26 genes per patient were affected based on SNVs only and all local variations including InDels, respectively. An even higher average number of affected genes were found in the analyzed parents (Figure 5.3A). In comparison, only 10 genes on average were found to contain potentially effective SNVs in the controls. In respect to this, the simple numeric excess of genes appears to favor the disease phenotype or the chance to give birth to affected children, respectively. Yet, the most extreme control individual showed effective SNVs in 25 genes, indicating that the specific type and pattern of mutation rather than the overall number is more important. Differences in the genetic background between TOF patients, parents and healthy controls were further delineated by a principal component analysis based on gene-wise SNV frequencies (data on InDels are not available in controls). Controls are characterized by SNVs in different genes than TOF patients and their parents (Figure 5.3B). Although patients and parents are more similar to each other than to the controls, they show a clear distinction studied separately (Figure 5.3C). Importantly, TOF patients are characterized by a common set of mutated genes, independent of whether they are members of the CHD families or represent unrelated sporadic cases.

A critical result of exome projects is the finding that a high number of potentially pathogenic variations can be observed in any healthy individual[286,287]. Most likely the combination of subsets of variations or their co-occurrence with external influences define the development of a disease state. Thus it is crucial to identify genetic variations relevant for the pathophysiology of a given disease. Using a permutation approach we assessed genes showing a significantly higher mutation rate (based on SNVs only to ensure comparability to the control dataset) in patients with TOF in comparison to healthy individuals (Danish controls). This resulted in 16 genes, which we defined

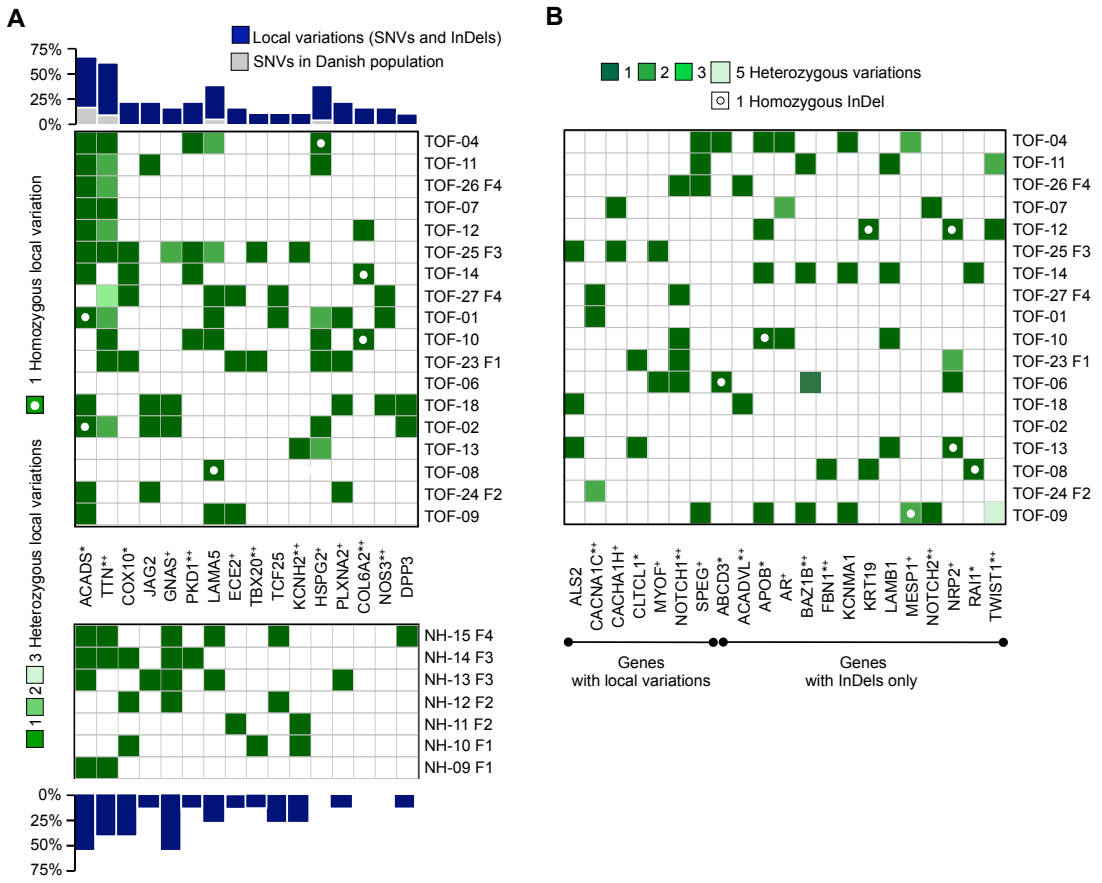

Figure 5.5: Distribution of TOF genes among cases. (A) Distribution of local variations (SNVs and InDels) found in the 16 significantly affected TOF genes (corrected p-value ≤ 0.05) in TOF patients (above) and healthy parents (below). Genes are ordered by significance from left to right. Gene-wise frequencies of local variants are represented by blue bars. Corresponding frequencies of SNVs in the control population (200 cases) are indicated in gray. (B) Distribution of local variations found in 21 potential TOF genes comprising genes not targeted in the controls (left) and genes with InDels only (right). (A and B) Familiar assignment is given after the sample identifier (F1 to F4). The number of local variations per gene is color-coded. Homozygous variations are additionally marked by a white dot. Genes marked with an asterisk have known associations with human disease affecting the heart, those marked with a cross show a cardiac phenotype when mutated or knocked out in mice.

as 'TOF genes'. First the observed ratio of each gene's mutation frequency (given as the total number of individuals that have at least one mutation in that gene) in TOF patients compared to healthy individuals (controls) was computed. A pseudocount of

$1x10^{-6}$ was added to every frequency to avoid zero counts. Afterwards, all individuals were randomly reassigned to individual mutation patterns to access a gene-wise distribution of mutation frequency ratios under random conditions. Following this approach, empirical p-values were derived by counting the number of random trials, where the found ratio exceeded the observed ratio, normalized by the number of trials. We used 100,000 and 10,000 random trials for gene- and SNV-wise significance, respectively, to ensure a high level of accuracy. Finally, only genes with Benjamini-Hochberg corrected (see Chapter 3.5) empirical p-value of less than or equal to 0.05 were defined as 'TOF genes'. These genes distinguish the TOF patients from the healthy controls, and moreover they are explanatory for the difference in the numerical excess of affected genes overall compared to the controls (Figure 5.4). Out of the 16 TOF genes, eight genes have known associations with human disease affecting the heart and ten genes show a cardiac phenotype when mutated or knocked out in mice (Figure 5.5A). Four of the TOF genes had not previously been associated with a heart phenotype. For further substantiation, we compared the mutation frequency of TOF genes to the central European population subgroup contained in the 1000 Genomes Project (exon Pilot dataset[286]). Using the same filtering criteria we found just one gene (PKD1) to contain potentially effective mutations.

Out of our 237 affected genes (based on local variations), 30 genes were not targeted in the controls. In addition, we found 67 genes harboring exclusively InDels (Supplementary Figure S5). The extraction of TOF-relevant genes out of this set (non-targeted and InDels only) is currently hindered by the lack of a control dataset. However, it is likely that additional genes out of this set will turn out to be relevant to TOF. Out of these 97 genes, we found 21 genes affected in at least two TOF patients, which we defined as potential TOF genes (Figure 5.5B). On average we found four TOF genes per patient (Figure 5.5A) with the majority of variations being SNVs and a minor proportion of InDels. For the patient TOF-06 no local variation could be found in the TOF genes. However, we found five variation in the potential TOF genes (Figure 5.5B) emphasizing their relevance to TOF.

To assess the significance of the mutation pattern that was found over the analyzed TOF patients, we compared the mutation frequency pattern found in the ten most significant genes against the control population. Based on the mutation frequencies in TOF, we defined three rules describing the observed (SNV-based) pattern: namely (A) two genes with a mutation frequency of at least 50%, (B) three genes with at least

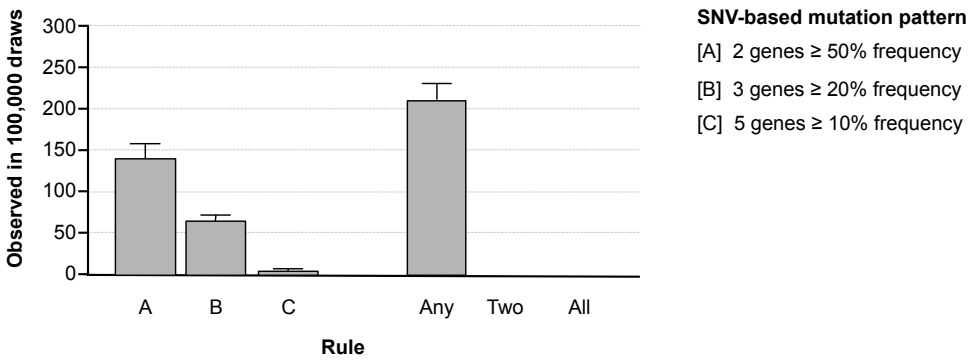

Figure 5.6: Statistical assessment of the mutation pattern of the ten most significant genes shows that the pattern is very unlikely to occur in a control population. Sets of 16 genes and 18 subjects of the control cohort were randomly drawn in comparison to the mutation pattern of the 10 most significant TOF genes. The bars indicate the average number of sets over ten times 10^5 draws which full-filled the defined rules (A) two genes showing a mutation frequency of at least 50%, (B) three genes showing a mutation frequency of at least 20%, which are not included in first, and (C) five genes showing a mutation frequency of at least 10%, which are not included in first or second. Additionally, the average number of sets is indicated which full-filled any, two or three of these rules.

20%, which are not included in first, and (C) five genes with at least 10% which are not included in first or second. An empirical p-value was derived from 10 times 10^5 randomly drawing groups choosing 16 genes (in accordance to the total number of defined TOF genes) and 18 individuals from the control population (in accordance to the number of analyzed TOF patients), a calculation of the resulting mutation frequencies and by comparison to the defined rules (average corrected empirical p-value for randomly drawn groups is given in Table 5.1). On average, the combinations fulfilling any individual rule were found 213 times and none of the cases exhibited two or all three rules (Figure 5.6). Choosing 37 genes (considering the significant and potential TOF genes) and 18 individuals, any individual rule was found 1,758 times, two rules were found six times and again, none of the cases exhibited all three rules. Thus, the observed mutation pattern in the TOF patients is very unlikely to occur in a healthy control subject. We further compared the individual mutation pattern of each of our TOF patients to controls and healthy parents and found no healthy individual showing exactly the same combination of affected genes.

To validate the pathological relevance of the variations observed in TOF genes, we

Rule	Average p-value
First	8.69e-04
Second	5.50e-04
Third	1.16e-04
Any	1.54e-03
Two	<1/100,000
All	<1/100,000

Table 5.1: Average empirical p-value for randomly drawn groups.

studied histological endomyocardial biopsy specimens (Figure 5.7A). For the patient TOF-08 only a single variation could be found. However, this variation is a homozygous deletion (5419delA, ENST00000252999) in the extracellular matrix gene laminin alpha 5 (LAMA5) and results in a frameshift leading to a truncated protein with loss of three essential protein domains (Figure 5.7B). The histological analysis in a respective right ventricular endomyocardial biopsy of the patient shows an abnormal configuration of myocyte alignment with branching fibers (Figure 5.7A).

The two most frequently affected genes (see Figure 5.5A) with an incidence of more than 50% of patients are mitochondrial short-chain specific acyl-CoA dehydrogenase (ACADS, also known as SCAD, Figure 5.7B) and titin (TTN). Two of the observed ACADS mutations that were observed, 625G>A (Gly209Ser, rs1799958) and 511C>T (Arg171Trp, rs1800556) are already known. For three cases carrying the 625G>A mutation (TOF-07, TOF-09 and TOF-11), we were able to study cardiac biopsies. Their histological analysis shows altered periodic acid schiff (PAS) staining, a feature which suggests a potential deficiency in mitochondrial function (Figure 5.7A). Titin is a key component of the sarcomere. All the TTN mutations observed in our TOF patients are heterozygous and occur in combination with other variations. For example they occur in combination with homozygous mutations of collagen VI alpha-2 (COL6A2, Figure 5.7B). Three of our TOF patients harbor mutations in the COL6A2 gene. Two are homozygous (TOF-10 and TOF-14 with 2096G>T [Gly699Val], ENST00000300527) and one is heterozygous with an allele frequency of 0.62 (TOF-12 with 1268C>T [Pro423Leu], ENST00000300527), which could also be validated by RNA-seq. Their potential impact can be observed by an increased assembly of collagen fibers in histological sections of respective cardiac biopsies (Figure 5.7A).

An overview of all genetic interactions of TOF genes together with their cellular function and localization is given in Figure 5.8. The latter were manually curated based

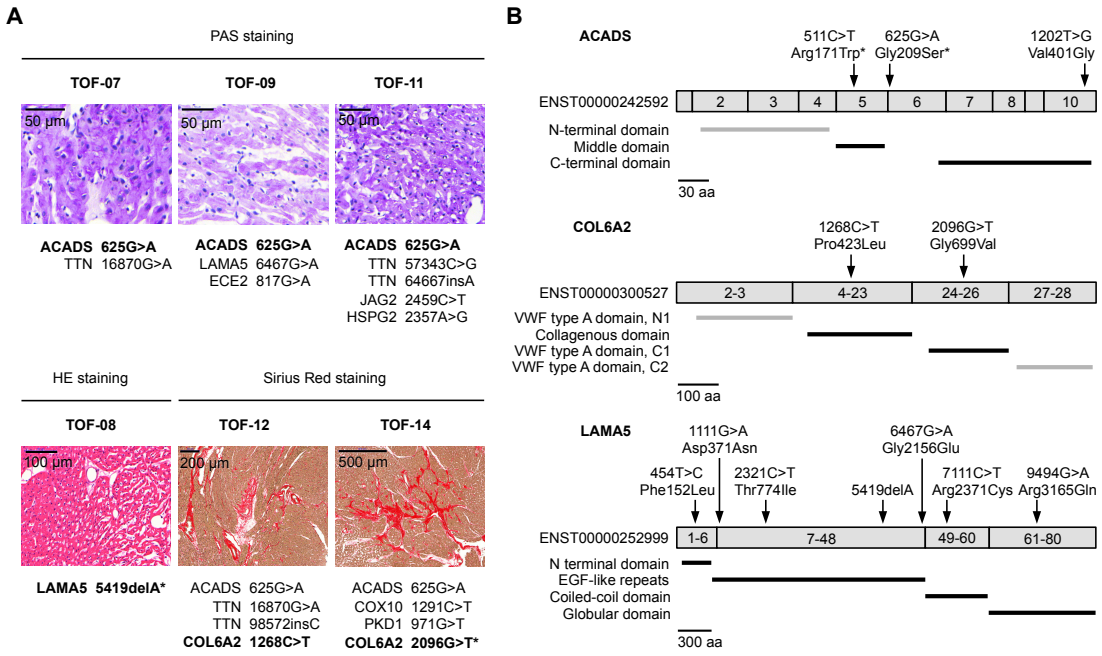

Figure 5.7: Functional consequences of mutations in TOF genes. (A) Histopathological assessment of right ventricular biopsies from selected TOF cases shows altered PAS staining (increase of PAS-positive granules), misalignment of the cardiac myocytes and increased interstitial fibrosis. Related mutations in TOF genes are listed for each subject. Those affecting genes relevant for the histological alterations are marked in bold and further depicted in (B). Variations marked with an asterisk are homozygous in the indicated case. (B) Location of the sequence variations in the protein structure of selected TOF genes. Coding exons are shown as grey boxes. Protein domains affected by variations are indicated as black lines, unaffected ones as grey lines. ACADS variations marked with an asterisk have been shown to reduce the protein's activity[290]. aa: amino acids.

on literature and the UniProt database (Supplementary Figure S6). The cellular localization for the TOF gene's proteins was first derived from the Swiss Prot annotation information (from the cellular component field) and for genes/proteins, which do not have cellular localization annotations, ConLoc and Proteome Analyst were used for the prediction of cellular localizations[291]. We analyzed cellular localizations of any genes showing SNVs and found no difference in the distribution between the TOF patients, their parents and healthy control individuals. However, an overrepresented proportion of TOF genes function in signal transduction pathways and are localized to the membrane, which highlights a role for TOF genes in regulatory signaling pathways.

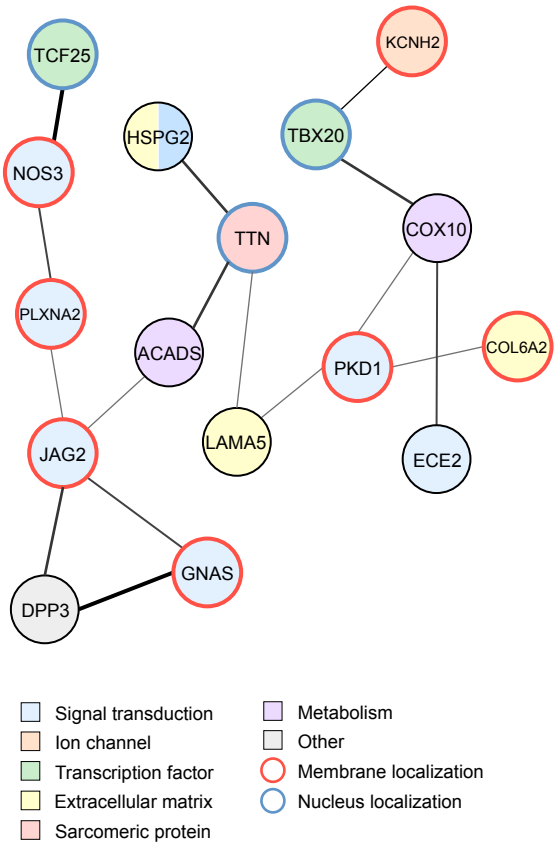

Figure 5.8: Genetic interaction network of TOF genes. Edges connect frequently pairwise mutated genes with a minimal normalized co-mutation frequency of 33%. Frequency is indicated by the line width. Manually curated functional categorizations of genes are color-coded. A red and blue border marks genes localized to the cell membrane and nucleus, respectively. The full list of functional characterizations and cellular localizations is given in Supplementary Figure S6.

To define genes of likely genetic interaction, the pairwise frequency of co-mutation as shown in Figure 5.8 was defined as the number of TOF patients showing mutations in both genes normalized by the number of patients, which showed a mutation of at least one of the two genes.

TOF is a developmental disorder and thus, causative genes have to be functional during embryonic development. To further verify the relevance of the identified TOF genes and potential TOF genes, a thorough literature analysis was performed, gathering data

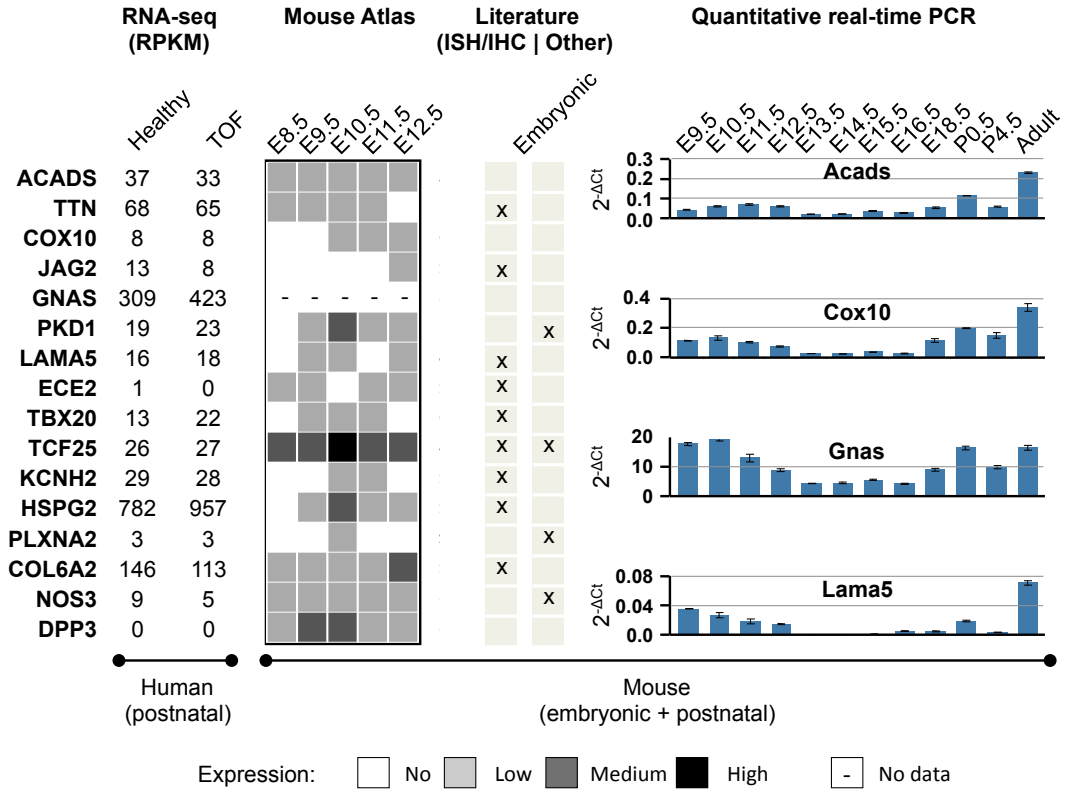

Figure 5.9: Expression of significant TOF genes in human and mouse. RNA-seq: average RPKM normalized expression levels in postnatal TOF and healthy unaffected individuals measured using mRNA-seq. Mouse Atlas: SAGE expression tag data of different developmental stages taken from Mouse Atlas of Gene Expression. If several different heart tissues have been measured, the maximum expression is shown. SAGE level is grouped into no (0), low (1-3), medium (4-7) and high (>7) expression. Literature: availability of published mRNA or protein expression data sets in mouse embryonic stages (E8.5 to E15.5) based on literature search including *in situ* hybridization (ISH)/immunohistochemistry (IHC) or other techniques (PCR, qPCR, Northern Blot and beta-galactosidase assay). The full list of data sets and corresponding publications can be found in Supplemental Figure S7. Quantitative real-time PCR: mRNA expression measurements in isolated mouse hearts of different embryonic and postnatal stages performed using qPCR. Expression values are normalized to housekeeping gene Hprt.

on mRNA and protein expression profiles based on techniques such as *in situ* hybridization or immunohistochemistry in human and mouse hearts at embryonic stages crucial for the development of TOF (week 3 to 10, E8.5 to E15.5, Figure 5.9 and Figure 5.10).

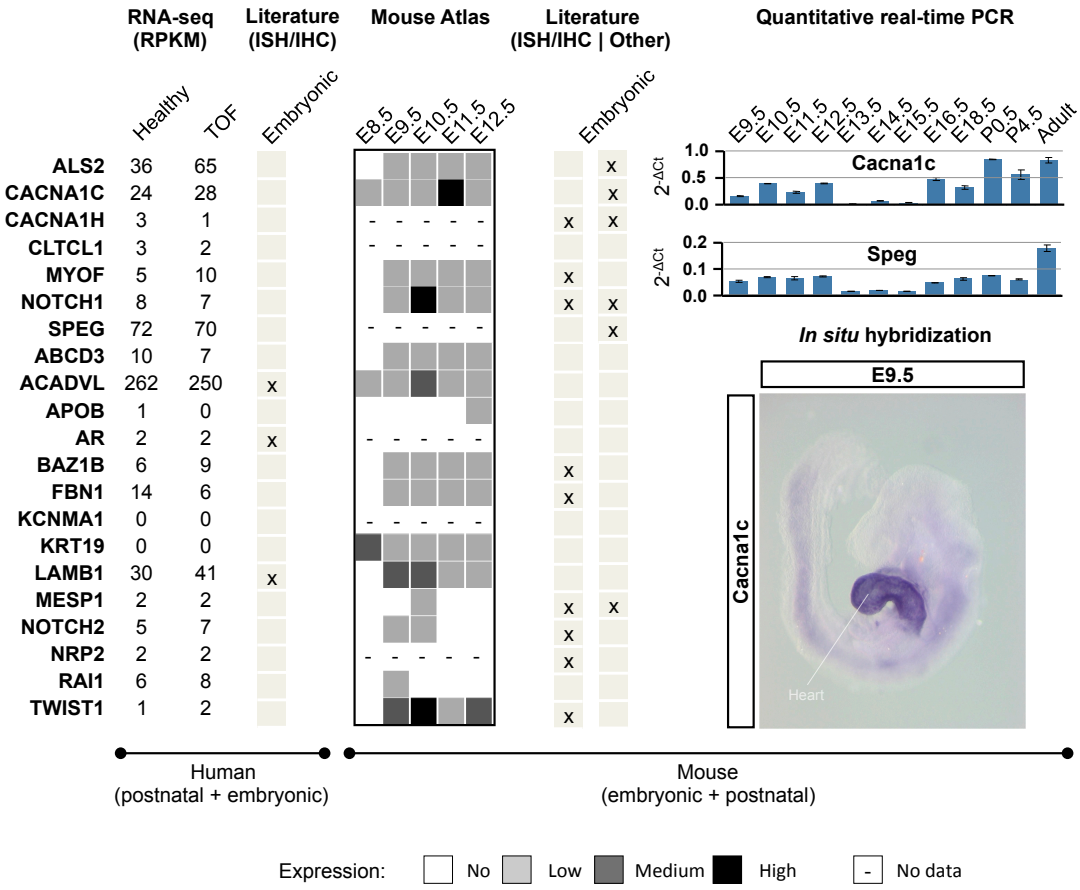

Figure 5.10: Expression of potential TOF genes in human and mouse. RNA-seq: average RPKM normalized expression levels in postnatal TOF and healthy unaffected individuals measured using mRNA-seq. Mouse Atlas: SAGE expression tag data of different developmental stages taken from Mouse Atlas of Gene Expression. If several different heart tissues have been measured, the maximum expression is shown. SAGE level is grouped into no (0), low (1-3), medium (4-7) and high (>7) expression. Literature: availability of published mRNA or protein expression data sets in human (week 3 to 10) and mouse (E8.5 to E15.5) embryonic stages based on literature search including *in situ* hybridization (ISH)/immunohistochemistry (IHC) or other techniques (PCR, qPCR, Northern Blot and beta-galactosidase assay). The full list of data sets and corresponding publications can be found in Supplemental Figure S7. Quantitative real-time PCR: mRNA expression measurements in isolated mouse hearts of different embryonic and postnatal stages performed using qPCR. Expression values are normalized to housekeeping gene Hprt. *In situ* hybridization: mRNA expression in E9.5 mouse embryo.

In addition, we evaluated embryonic gene expression profiles of the TOF genes using SAGE data from the Mouse Atlas of Gene Expression project[292]. This combined approach revealed only six genes (one TOF gene and five potential TOF genes) that were not already known to be expressed during the embryonic development of the mouse heart, and one (CLTCL; potential TOF gene) lacks a mouse homolog altogether. To further extend these data, quantitative real-time PCR (qRT-PCR) was performed in our group for six genes in mouse hearts at the developmental stages E9.5 to E18.5, postnatal at P0.5 and P4.5 as well as at adulthood (Figure 5.9 and Figure 5.10). Strikingly, all of these genes show an embryonic expression at the crucial developmental phase and all have a biphasic profile with continued expression postnatal and at adulthood. The cardiac expression of Cacna1c during development was further demonstrated using whole mount *in situ* hybridization at E9.5 mouse embryos (Figure 5.10). Based on gene expression profiles obtained by RNA-seq (see the following Chapter 5.3), we found the majority of genes being expressed (RPKM>1; gene expression analysis is given in the following Chapter 5.3) in the human right ventricle of TOF patients as well as in normal adult hearts (Figure 5.9 and Figure 5.10). As the RPKM value as measured by RNA-seq should be proportional to the average mRNA numbers per cell, genes can be defined as lowly expressed (RPKM$\leq$1) or highly expressed genes (RPKM>1), respectively[293].

Finally, we were interested in the segregation of identified mutations in TOF genes within our studied families. We observed a combination of novel and inherited mutations in these genes in the affected family members (Supplementary Figure S8), which is in line with a non-Mendelian inheritance. The finding that a certain number of mutations are inherited underlines our observation of the general numeric excess of mutations in parents compared to other healthy individuals.

5.3 Gene Expression Analysis

For expression analysis RNA profiles were gathered from right ventricle of 22 patients with TOF as well as from left and right ventricle (LV and RV, respectively) of four healthy unaffected individuals. Deep sequencing of the mRNA libraries resulted in ∼19,224,000 single-end reads (36 bp) per sample on average (see Chapter 2.2.4). The reads were mapped to the human reference genome (NCBI v36.1, hg18) using RazerS[174] allowing at most 10 equally-best hits and two mismatches (no InDels) per read. On average, ∼14,736,000 reads per sample were mapped to the whole human reference genome. Approximately 9,431,000 reads (64%) per sample could be mapped to unique genomic locations and ∼5,304,000 reads (36%) matched to multiple regions (2-10 genomic locations). Multi-matched reads were proportionately assigned to each of their mapping locations using the MuMRescueLite[212] approach with a window size of 200 bp. The distribution of the read counts over all patients and healthy individuals after sequencing and mapping is given in Figure 5.11. Reads that were found in unique or multiple positions in the human genome were assigned to genes if their mapped location is inside of exon boundaries as defined by ENSEMBL[207] (v54). Finally, the number of reads that were fully included in exons was counted. On average 79% (±4%) of the mapped reads could be assigned to known exons. A high percentage of the mapped reads (25±4%) was assigned to exons located on the mitochondrial chromosome. This is in line with the fact that the heart muscle is rich in mitochondria, which are responsible for the energy metabolism of the cell. The mitochondrial genome encodes several subunits of the mitochondrial respiratory chain such as cytochrome c oxidase and NADH dehydrogenase. To further assign unmapped reads, a gene-wise splice junction sequence library was produced from pairwise connection of exon sequences corresponding to all known 5' to 3' splice junctions (supported by the analysis of aligned EST and cDNA sequences). Over all samples 21.8% of the previously unmapped reads were mapped on average to the set of known splice junction sequences using RazerS allowing at most two mismatches (no InDels) and only unique best matches.

Read count normalization for mRNAs was performed using the RNA composition adjustment by trimmed mean of M-values (TMM) and quantile-to-quantile count adjustment implemented in the edgeR[236] package (see Chapter 3.4.2). For quality assessment manual inspection of multi-dimensional scaling (MDS) plots and existence of pile-up effects were performed. First, a plot showing the sample relations based on multidimensional scaling was produced (Figure 5.12). The distance between each pair of samples

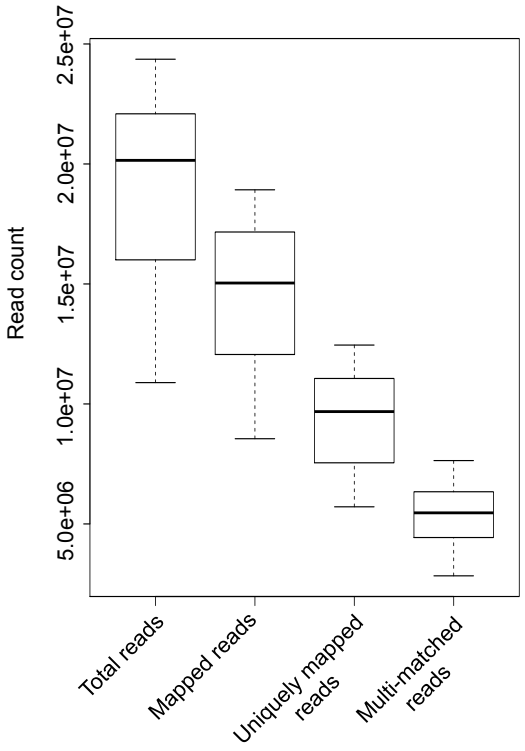

Figure 5.11: Distribution of read counts over 22 TOF patients and left as well right ventricle of four healthy individuals after mRNA sequencing and mapping to the human reference genome.

was calculated based on the square root of the common dispersion for the top 5,000 genes which best distinguish that pair of samples. These genes were selected according to the tag-wise dispersion of all the samples. From this plot, four samples (TOF-11, TOF-14, TOF-18 and TOF-19) were identified as outliers due to their large distance to the other TOF samples in the first dimension. For the healthy individuals two samples from the left (NH-07) and right (NH-08) ventricle from one individual appear to be separated from the other normal heart samples in the second dimension. However, the distances between the normal heart samples are relatively small, thus we did not treat these samples as outliers. Second, we examined the number of duplicated sequencing reads before and after read mapping to identify possible mapping or PCR problems. On average 52% ($\pm$8%) of the sequencing reads over all samples are represented by unique sequences (i.e. one read represents one sequence but this sequence can be represented by n other reads) and 87% ($\pm$2.9%) of these unique sequences are represented by one read and 12% ($\pm$2.6%) by 2-10 reads (Supplementary Figure S9). After read map-

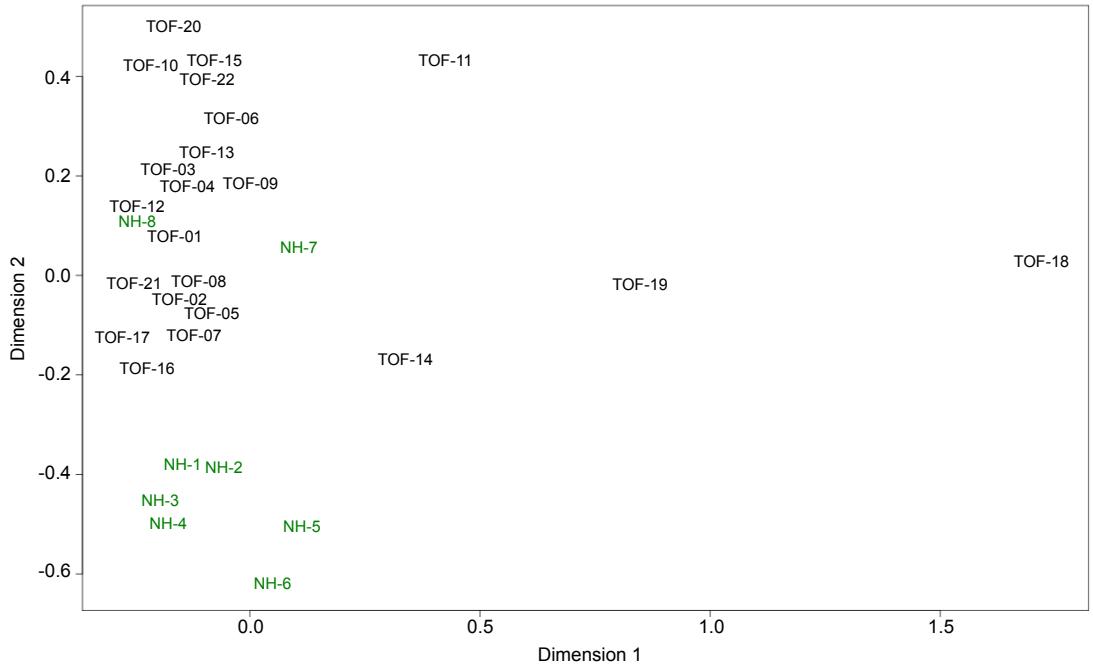

Figure 5.12: Multidimensional scaling (MDS) plot based on gene expression levels (top 5,000 genes which best distinguish that pair of samples) for the RNA-seq data obtained from mRNA libraries of patients with Tetralogy of Fallot (TOF) and healthy unaffected individuals (normal heart, NH).

ping we checked for the uniquely mapped reads the total number of perfectly identical start/end sites and found two samples (TOF-18 and TOF-19) with a significantly lower number of start/end sites represented by one read and a significantly higher number of start/end sites represented by more than 1,000 reads (Supplementary Figure S10). Interestingly, the mRNA library of TOF-18 had to be sequenced three times to obtain the necessary sequencing quality and output. In addition, the samples TOF-11 and TOF-14 also showed pile-up effects in the unique read mapping process, in particular for start/end sites represented by 101-1000 reads (Supplementary Figure S10). In summary, all four samples identified as outliers in the MDS plot (TOF-11, TOF-14, TOF-18 and TOF-19, Figure 5.12) were removed from further analysis.

To define differential expression between affected and healthy individuals, an significance test based on the negative binomial distribution for tag-wise dispersion (see Chapter 3.4.3) also implemented in the edgeR package was applied to genes with a minimal read count of 100 over all analyzed samples. Since it is not possible to achieve statistical significance with very low total counts, we discarded those genes, thereby

	TOF vs. RV	TOF vs. LV	RV vs. LV
Sig. diff. expressed genes	1,514	1,788	182
- upregulated genes	633 (42%)	864 (48%)	89 (48%)
- downregulated genes	881 (58%)	924 (52%)	93 (52%)
Sig. diff. expressed transcripts	1,765	2,036	208
(corresponding genes)	(1,390)	(1,607)	(208)
- upregulated transcripts	616 (35%)	818 (40%)	107 (51%)
- downregulated transcripts	1,149 (65%)	1,218 (60%)	101 (49%)

Table 5.2: Significantly differentially expressed genes and transcripts with p-value < 0.05 after adjustment for multiple testing in 18 patients with Tetralogy of Fallot (TOF) versus left (LV) and right (RV) ventricle of four healthy individuals as well as RV versus LV.

excluding lowly expressed mRNAs that only contribute to noise. In total, 26,522 genes were found to be expressed with at least one exonic or junction read over all analyzed mRNA-seq samples. Almost half of these genes (48.6%) are lowly expressed according to their RPKM value, i.e. RPKM$\leq$1 on average over all analyzed samples. After discarding the lowly expressed genes based on the raw read count level, 17,184 genes were used for differential gene expression analysis. However, based on RPKM values instead of read count levels there were still lowly expressed genes (21.5%) but the median RPKM value over the retained genes could be increased from 1.1 to 3.9. Moreover, the RPKM value for the lower quantile could be increased from 0.1 to 1.2 (i.e. from lowly to highly expressed according to Hebenstreit *et al.*[293]) and for the higher quantile from 6.2 to 11.6. Using the tag-wise dispersions we found 1,514 genes (8.8% of all analyzed genes) to be significantly differentially expressed between right ventricle of 18 TOF patients and four healthy individuals (RV) with a Benjamini-Hochberg corrected p-value (see Chapter 3.5) of less than 0.05. Of these genes, 881 (58%) were upregulated in TOF versus RV and 633 (42%) were downregulated. In addition, we performed differential gene expression analysis between TOF and left ventricle of the four healthy individuals (LV) as well as RV versus LV (Table 5.2).

Further, we analyzed the gene expression similarity between TOF and RV of all expressed genes measured by normalized Euclidean distance. The level of expression similarity is high within the individual groups ($\sim$0.5 for RV and $\sim$0.7 for TOF). However, the similarity between the two groups is low ($\sim$0.35), indicating a commonly changed expression profile in TOF patients (Figure 5.13). We analyzed this in more detail including also LV. The gene expression similarity was again measured by normalized Euclidean distance and repeatedly, TOF against TOF was most similar. In

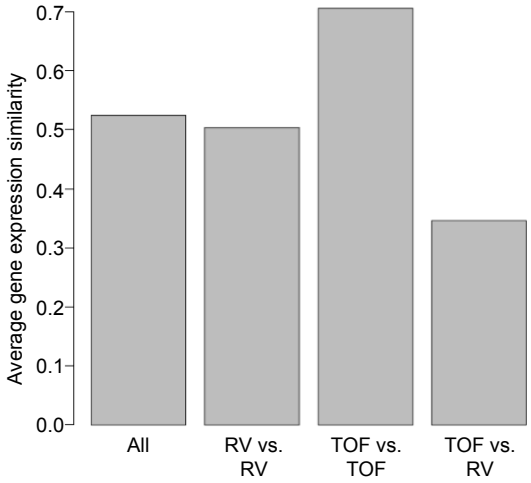

Figure 5.13: Average gene expression similarity measured by normalized Euclidean distance between TOF patients and right ventricle of healthy individuals (RV).

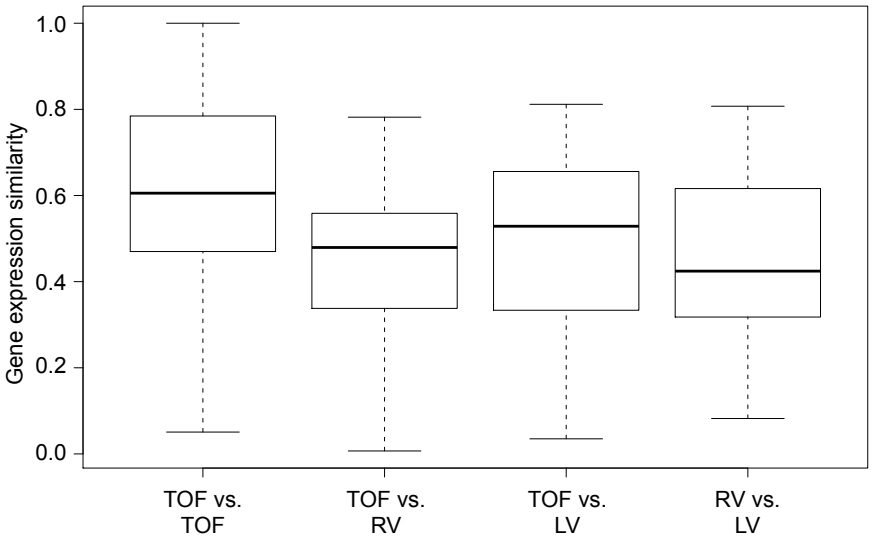

Figure 5.14: Boxplots for pairwise gene expression similarity measured by normalized Euclidean distance over all individuals in either TOF, healthy right ventricle (RV), healthy left ventricle (LV) or between these groups.

addition, TOF against LV has a higher similarity level than TOF against RV (Figure 5.14).

Beside genes, we also computed the differential expression of transcripts, whose read

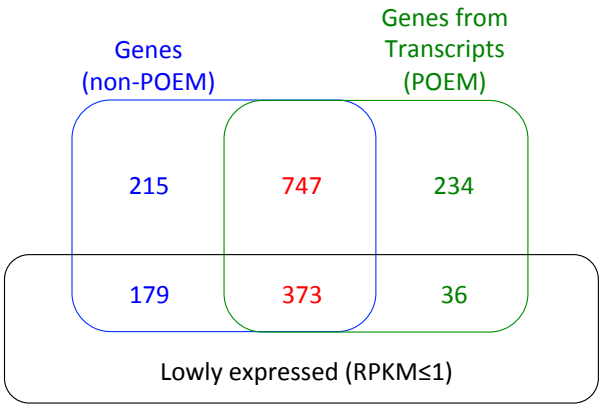

Figure 5.15: Overlap of significantly differentially expressed genes (in total 1,514 non-POEM genes) and genes (in total 1,390) representing significantly differentially expressed transcripts (POEM-genes) in TOF patients versus RV as well as their overlap to lowly expressed genes (RPKM≤1) over all analyzed samples.

counts were adjusted using the POEM method described in Chapter 3.3.1.1. As for the genes, the read counts were normalized using the TMM normalization method followed by quantile-to-quantile count adjustment (see Chapter 3.4.2). Again, the corresponding MDS plot was evaluated for quality assessment, leading to the exclusion of three samples (TOF-14, TOF-18 and TOF-19) for further analysis. The negative binomial distribution test for tag-wise dispersion (see Chapter 3.4.3) was applied to transcripts with a minimal read count of 100 over all analyzed samples to define differential expression between TOF and healthy individuals. Finally, 1,765 transcripts were found to be significantly differentially expressed between TOF and RV with a Benjamini-Hocherg corrected p-value (see Chapter 3.5) of less than 0.05 (Table 5.2). These transcripts refer to 1,390 genes, which we called 'POEM genes'. We compared these POEM genes with the 1,514 significantly differentially expressed genes and as expected found a high overlap of 81% (Figure 5.15). 45% of the non-overlapping non-POEM genes and only 13% of the non-overlapping POEM genes are lowly expressed. This suggests that differential expression of genes with low expression levels is detected more frequently by the gene-based approach than by the isoform-based approach, which is likely depended on lower read counts. To detect differential expression of genes with high expression (RPKM>1) both the gene- and isoform-based approach are reliable. However, there are non-overlapping highly expressed genes, but they are mostly borderline significant. For example, 62% of the 215 non-overlapping non-POEM genes with p-value less than 0.05 overlap to POEM-genes with increased FDR (10%). In summary, the significantly

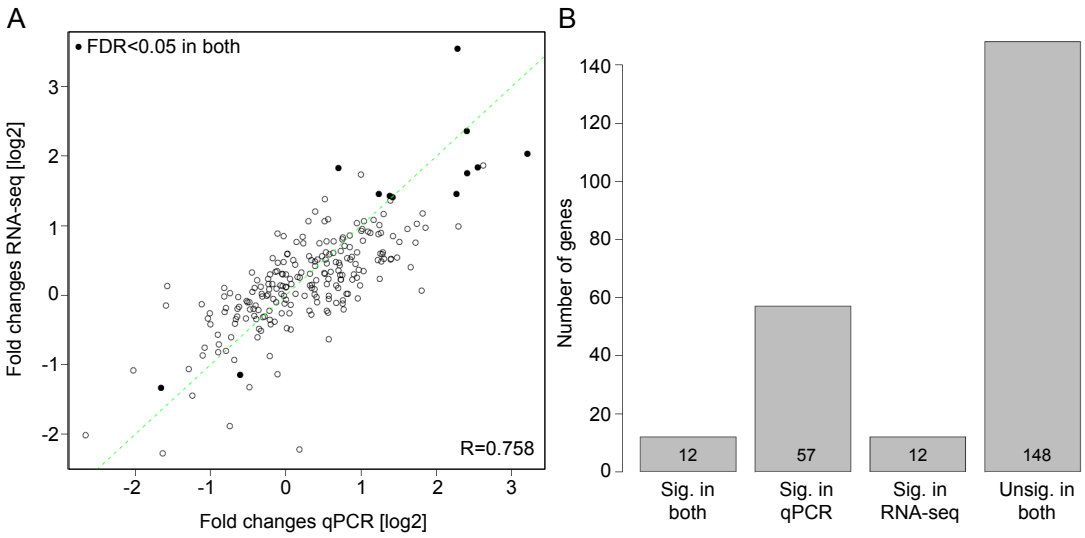

Figure 5.16: (A) Comparison of fold changes between RNA-seq and qPCR data measured by the Lightcycler 1536 system and (B) their number of genes significantly differentially expressed between TOF patients and right ventricle of healthy individuals. To compare the qPCR results to the mRNA-seq data the number of genes was reduced to those measured in both experiments.

differentially expressed genes quantified by the gene-based approach are used for further analysis.

We further compared differential expression in TOF against RV measured by RNA-seq to data generated in a previous study using the Roche LightCycler 1536 system for high-throughput quantitative real-time PCR[294]. Briefly, using the Lightcycler 245 genes were measured in triplicates in the same human heart samples from TOF patients and healthy individuals (RV) as in our RNA-seq data. The average expression value was calculated for each set of triplicates after manual outliers removal. The expression levels were further normalized using the geometric mean of three housekeeping genes (HPRT, B2M and GAPDH). Finally, we calculated differential expression for the normalized expression values between TOF patients and healthy individuals. We found 70 significantly differentially expressed genes with a Benjamini-Hocherg corrected p-value (see Chapter 3.5) of less than 0.05 using a t-test. In contrast to mRNA-seq, most of the genes (94%) measured by qPCR are upregulated (in total 66 genes) and only few are downregulated (in total 4 genes). To compare the results to the mRNA-seq data we reduced the number of genes to those measured in both experiments. In general, the measured fold changes between mRNA-seq and qPCR data are well correlated (Pearson

correlation coefficient of ∼0.8, Figure 5.16A). However, the actual sets of genes that were commonly found to be differentially expressed had a rather modest overlap with a high number of genes that were only differentially expressed in qPCR but not in mRNA-seq (Figure 5.16B).

Finally, after mapping all previously unmapped reads to known splice junction sequences, we mapped the remaining reads to a set of candidate novel splice junctions, which correspond to all hypothetical additional 5' to 3' pairings of splice sites in the same set of genes. Over all samples 45,430 candidate novel splice junctions with at least one mapped sequencing read could be identified. Increasing the number of junction reads to at least 10 over all samples resulted in 4,278 previously unknown splice junctions and 1,175 novel splice junctions with at least 50 mapped reads. Searching for novel splice junctions with more than 10 mapped reads on average over all samples in either TOF, RV or in both, we found alternative splicing events in 216 genes representing 279 potential novel splice junctions. Among these genes are several sarcomeric genes such as cardiac troponin T (TNNT2), cardiac troponin I (TNNI1) and myosin heavy chain 7 (MYH7). It has been shown that associated changes in mRNA splicing of these three genes were significantly altered in patients with ischemic cardiomyopathy, dilated cardiomyopathy and aortic stenosis[295]. We selected five identified candidate novel splice sites in the genes TNNI1, MYL7, PPARG and PDLIM3 (Table 5.3) for reverse transcription PCR (RT-PCR) validation in one healthy individual (NH-04) and three TOF patients (TOF-03, TOF-06 and TOF-11).

In detail, TNNI1 is expressed in cardiac and skeletal muscle in early development but restricted to slow twitch skeletal muscle fibers in adults[296]. The candidate novel splice site in TNNI1 located 4 amino acids downstream of the start codon generates a transcript which is composed of one incomplete (missing the 5'UTR) and one well annotated transcript (Figure 5.17). The splice site generates a frameshift that leads to an altered amino acid sequence and to a termination of the protein 16 amino acids after the splice site, resulting in a non-functional protein. In line with our mRNA-seq data, the expression of the alternative transcript measured by RT-PCR was stronger in the TOF patients than in the healthy individual.

MYH7 encodes the cardiac muscle beta (or slow) isoform of myosin and changes in the relative abundance of MYH7 correlate with the contractile velocity of cardiac muscle[297]. In addition, we found an upregulation of the atrial myosin regulatory light chain in the hypertrophic ventricle of our TOF patients. The identified novel splice site in

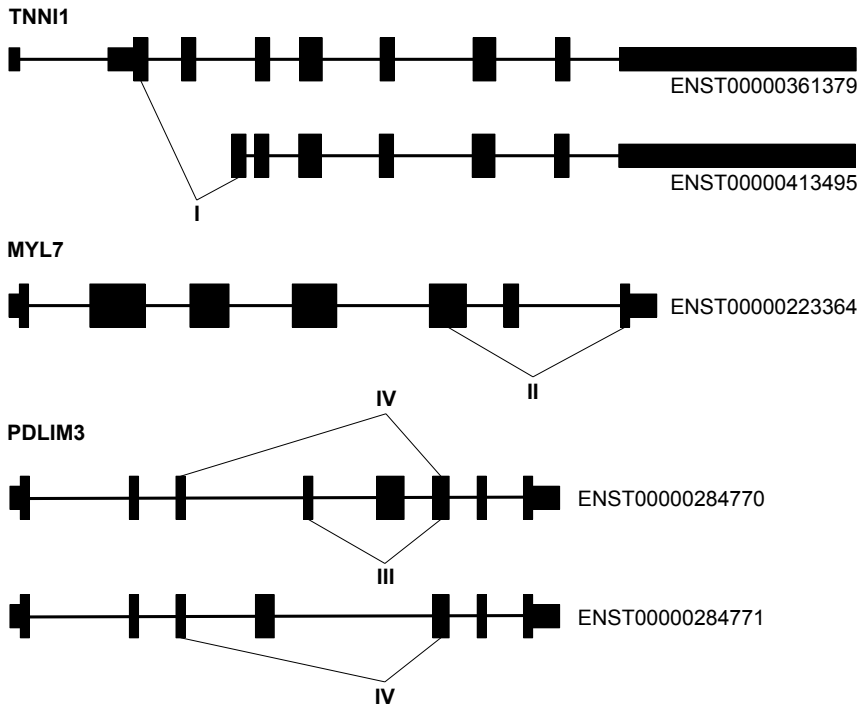

Figure 5.17: Schematic representation of candidate novel splice junctions in the genes TNNI1, MYH7 and PDLIM3 (based on Ensembl v59 and not drawn to scale) identified by RNA-seq in either TOF patients, healthy individuals or in both. Details for the individuals splice junctions (I-IV) are given in Table 5.3.

MYL7 removes an exon from a well annotated transcript (Figure 5.17). The splice site could be detected in human heart cDNA (validated by sequencing of the PCR products). Moreover, the RT-PCR also showed a weak upregulation of MYL7 in TOF patients and a slightly stronger expression of the novel splice site in TOF compared to normal heart. This splice site leads to a truncation of the second EF hand domain and in addition, a frameshift generates a stop codon very shortly (6 aa) after the novel splice site. The PDZ and LIM domain protein 3 (PDLIM3, also known as ALP) is involved in cytoskeletal assembly and has two major isoforms. Both isoforms were measured by RNA-seq in our healthy individuals with 64% (ENST00000284771) and 31% (ENST00000284770) of the transcripts according to our POEM estimations. Each transcript has a different tissue-specific ZM motif. The two novel splice sites remove exons from the two major isoforms (Figure 5.17). The expression of both splice sites as well as the known transcripts could be validated by sequencing of the PCR products. The RT-PCR showed the downregulation of PDLIM3 in TOF patients. Moreover, a

ID	Gene	Ensembl exon-exon junctions	Mean junction reads in RV	Mean junction reads in TOF	Mean RPKM in RV	Mean RPKM in TOF	P-value TOF vs. RV
I	TNNI1	ENSE00001510261-ENSE00001350131	0.3	17.2	4	167	5e-14
II	MYL7	ENSE00001176203-ENSE00000680788	45.3	143.2	1,187	3,135	0.02
III	PDLIM3	ENSE00002526712-ENSE00002536022	13.5	1.4	124	30	3e-07
IV	PDLIM3	ENSE00002526712-ENSE00002464627	0.5	0.5	124	30	3e-07
-	PPARG	ENSE00001527052-ENSE00001527016	9.3	10.8	3.2	1.2	0.004

Table 5.3: Candidate novel splice junctions in the genes TNNI1, MYH7, PDLIM3 and PPARG identified by RNA-seq in TOF patients and right ventricle of healthy individuals (RV). Except for PPARG the Ensembl exon IDs are based on release v65 (hg19). For PPARG the transcript ENST00000397003 with exon ENSE00001527016 was removed from Ensembl v65, therefore the IDs for Ensembl v54 (hg18) are provided. The splice junction I-IV could be validated by RT-PCR and correspond to identifiers used in Figure 5.17.

shift in the ratio of the two known major isoforms from 2:1 in the healthy individual to 15:1 in the TOF patients was observed as well as a higher expression of the novel spice sites in the healthy individual compared to TOF. The splice site III (Figure 5.17 and Table 5.3) leads to a deletion of the ZASP domain in the major transcript. While the splice site III generates an intact protein, the splice site IV (Figure 5.17 and Table 5.3) causes a frameshift that leads to the termination of the protein (21 aa after the novel splice site). Notably, the ZASP domain is important for binding the rod region of alpha-actinin[298].

The peroxisome proliferator-activated receptor gamma (PPARG) is a nuclear receptor that regulates adipocyte differentiation. It has been implicated in the pathology of numerous diseases including obesity, diabetes, atherosclerosis and cancer. The candidate novel splice site joins two incompletely annotated transcripts, of which one (ENST00000397003) is even removed from the current Ensembl version 65. The splice site could not be validated by RT-PCR, but the downregulation of the gene in TOF patients could be shown. In summary, four out of five selected candidate novel splice sites could be validated by RT-PCR, with the quality of the transcript annotation being a possible indicator for the validation success.

5.4 MicroRNA Profiling

Deep sequencing of small RNA libraries from 22 TOF patients and left (LV) as well as right ventricle (RV) of four healthy unaffected individuals produced approximately 450 million raw reads of 36 bp in length (on average ∼15 million reads per sample; see Chapter 2.2.4). To prevent multiple mapping of identical small RNA sequences, redundancy was removed meaning that reads with an identical sequence were represented with a single entry storing the number of sequence counts. This yielded ∼170 million non-redundant (unique) read sequences (on average ∼5.6 million per sample). The unique read sequences were mapped to the human reference genome (NCBI v36.1; hg18) using MicroRazerS (Chapter 3.1). The parameters were set as follows: -m 20 (maximum number of best matches), -pa (purge ambiguous reads having more than 20 equally-best hits) and -sL 18 (seed length). Searching for miRNAs having a length of 19-25 nt, we found a minimal length of 18 nt to be good seeds to start the read mapping process. In addition, we allowed to map reads with at most one error in the seed sequence to be robust towards possible sequencing errors and sequence variations. On average ∼4,754,960 unique sequences per sample could be mapped to the human reference genome representing 91% of the total reads (∼13,595,423) per sample. Multi-matched reads were proportionally assigned to their loci. Using annotations from miRBase[43] database (v14), 53% of all mapped reads (on average ∼7,254,323 reads per sample) could be assigned to known mature miRNA sequences, which we called miRNA reads (Figure 5.18). Searching for reads overlapping with known precursor miRNA sequences we found only few additionally mapped reads (on average ∼7,388,436 reads per sample) indicating that almost all miRNA read sequences are products of functional miRNA strands (Supplementary Figure S11).

Total reads	449,996,875	
Mapped reads	407,862,701	
- miRNAs	221,653,090	54.3%
- other small RNAs	5,293,886	1.3%
- mRNA	8,886,443	2.2%
- repeats	49,889,283	12.2%
- unknown	122,139,999	29.9%

Table 5.4: Total number of reads over all samples after RNA sequencing and mapping to the human reference genome and their distribution to known miRNAs, other small non-coding RNAs, mRNA sequences and genomic repeats.

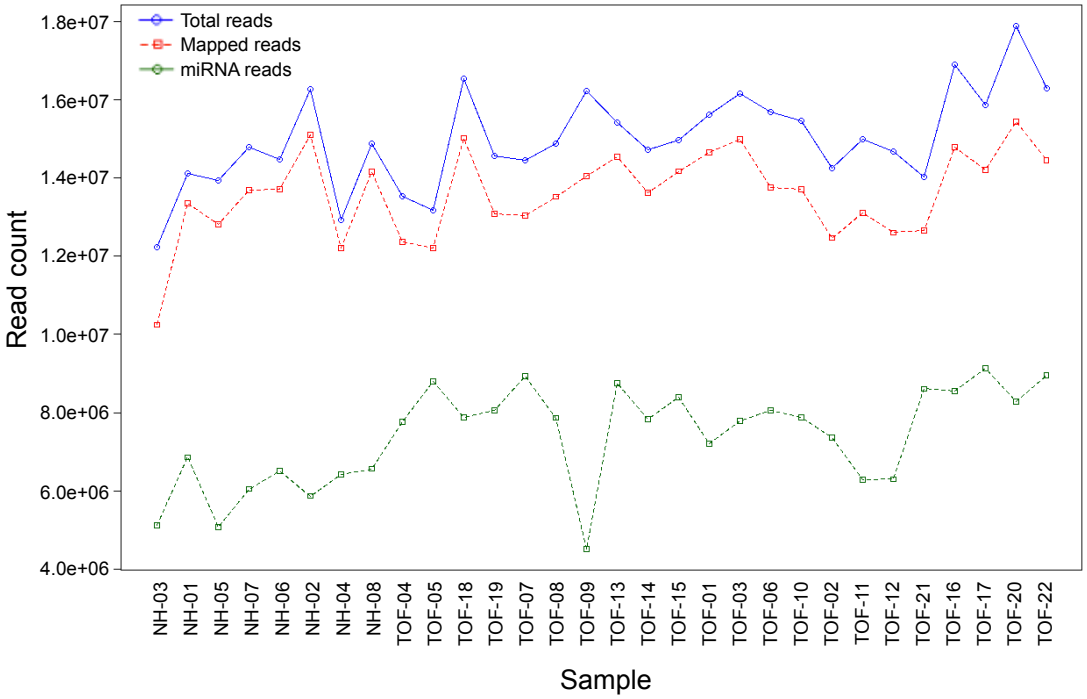

Figure 5.18: Small RNA read counts for TOF patients and healthy individuals (normal heart, NH) after sequencing, mapping and annotation.

The mapped small RNA read sequences show a bimodal length distribution with two distinct peaks representing miRNAs as well as other non-coding RNAs (Figure 5.19A). However, after annotation we found a length distribution representative for miRNA sequences, i.e. 18-25 nucleotides with a single peak near the average mapped read length of 22.3 nucleotides (Figure 5.19B).

After annotation to known human miRNAs, we assigned the remaining mapped reads to other known non-coding RNAs, mRNA sequences and genomic repeats using annotations from the UCSC[219] database (Figure 5.20A and Table 5.4). On average over all analyzed samples we found that the most abundant classes of non-coding RNAs except miRNAs are rRNAs and tRNAs (Figure 5.20B). However, we observed only a low number of other small non-coding RNAs (1.3%) and mRNA sequences (2.2%) indicating an accurate library preparation and low contamination over all small RNA-seq libraries. The relatively high number of reads assigned to genomic repeats could be explained by ambiguously mapped reads due to the relatively loose criteria of a 18 bp seed and the number of equal-best hits (at most 20) in the read mapping process.

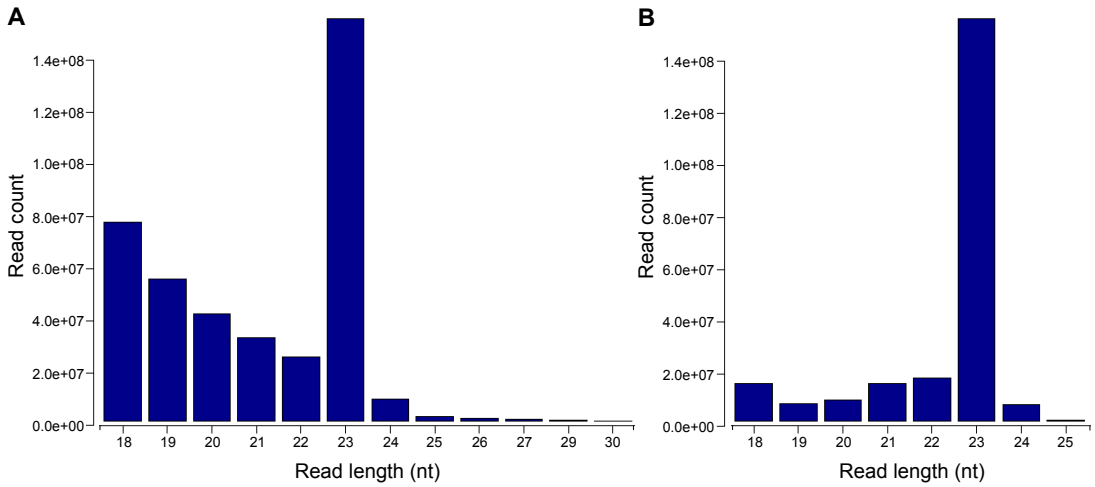

Figure 5.19: Lengths of (A) mapped small RNA read sequences and (B) annotated miRNA read sequences over all analyzed samples.

Finally, small RNA-seq revealed on average 396 expressed miRNAs per sample representing 450 loci. A higher number of expressed miRNAs was found in the TOF patients (on average 413 miRNAs per sample representing 463 loci) compared to the healthy individuals (on average 363 miRNA per sample representing 407 loci). For miRNA read count normalization we used the TMM normalization method followed by quantile-to-quantile count adjustment (see Chapter 3.4.2). After miRNA quantification and read

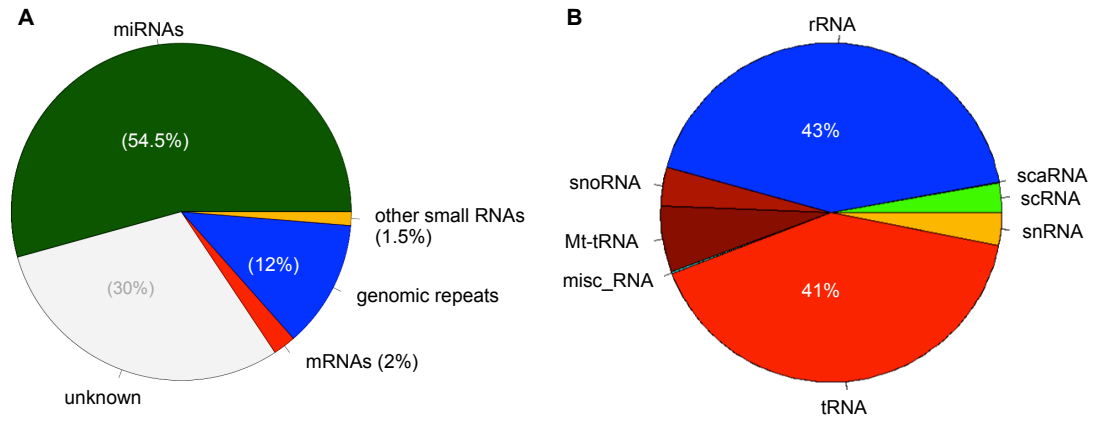

Figure 5.20: (A) Annotation of read sequences over all analyzed small RNA-seq samples and (B) annotations of small non-coding RNAs except miRNAs.

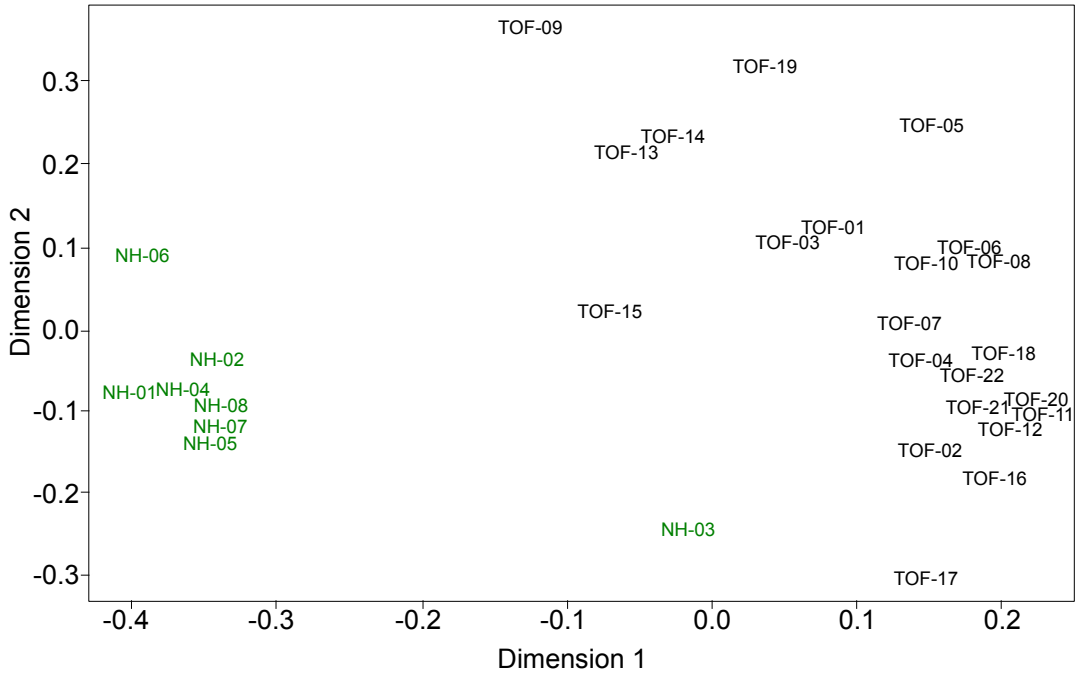

Figure 5.21: Multidimensional scaling (MDS) plot based on miRNA expression levels for the miRNA-seq data obtained from small RNA libraries of patients with Tetralogy of Fallot (TOF) and healthy unaffected individuals (normal heart, NH).

count normalization we assessed the sample relations based on multidimensional scaling, resulting in the exclusion of two samples for further analysis. From the MDS plot (Figure 5.21) we identified the normal heart sample NH-03 as an outlier because it was clearly separated from the other normal heart samples in the first dimension. For the TOF patients the sample TOF-09 may be identified as an outlier but the distance between this sample and the other TOF samples is relatively small in both dimension (Figure 5.21). Thus, we additionally did a classical principal component analysis which identified this sample as a clear outlier. In addition, the TOF-09 sample had a significantly lower number of annotated miRNA reads than the other TOF samples (Figure 5.18) and moreover, a much higher number of reads was assigned to known mRNAs as well as genomic repeats. In summary, besides NH-03 the sample TOF-09 was also removed from further analysis.

To further analyze the miRNA expression profiles we calculated the pairwise miRNA expression similarity measured by Euclidean distance over all individuals in either TOF, healthy right ventricle (RV), healthy left ventricle (LV) or between these groups (Fig-

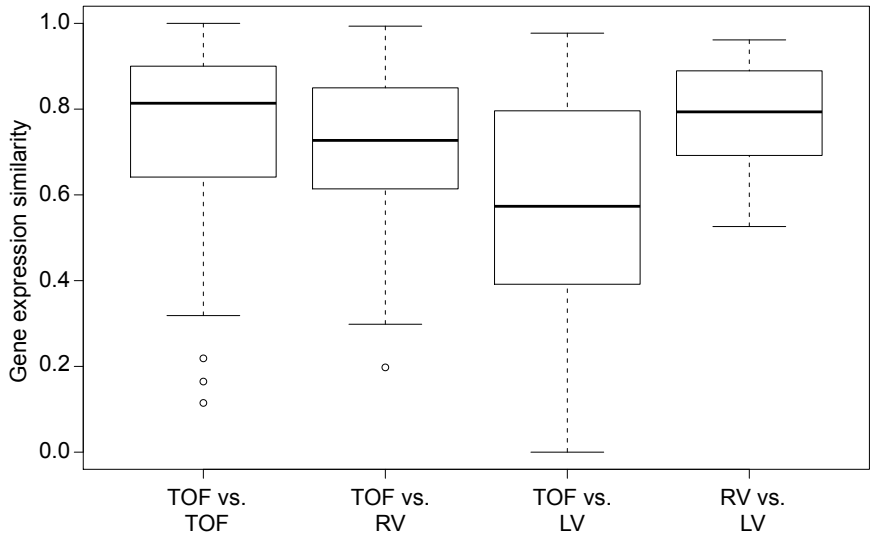

Figure 5.22: Boxplots for pairwise miRNA expression similarity measured by Euclidean distance over all individuals in either TOF, healthy right ventricle (RV), healthy left ventricle (LV) or between these groups.

ure 5.22). We found the TOF patients to be most similar to each other. However, compared to the gene expression similarity (Figure 5.14) TOF against LV loses the similarity. Moreover, we considered the average correlation in expression over all TOF patients and healthy individuals between miRNAs residing in the same family. As expected, miRNAs belonging to the same family show a high positive correlation (average Pearson correlation coefficient of 0.64 ± 0.2 over 54 miRNA families).

In summary, we found 626 expressed miRNAs with at least one read over all samples. To define differential expression between healthy and affected individuals, the negative binomial distribution test for tag-wise dispersion (see Chapter 3.4.3) was applied to miRNAs with a minimal tag count of more than 100 over all analyzed samples. The analysis revealed 103 significantly differentially expressed miRNAs (33.1% of all analyzed miRNAs) between TOF and healthy individuals (RV) with a Benjamini-Hochberg corrected p-value see Chapter 3.5) of less than 0.05 (Supplementary Table S6). Most of these miRNAs (in total 93) were upregulated in TOF versus RV including several heart- and muscle-relevant miRNAs (e.g. let-7b/c, miR-221, mirR-222, miR-378, miR-10a, miR-127, miR-30b and miR-15b). Only few miRNAs (in total 10) were downregulated in TOF patients including the muscle-specific miR-133b as well as miR-29b/c, which are involved in the control of cardiac fibrosis via mRNA repression of collagens, fibrillins

and elastin[299]. The downregulation of miR-29 induces the expression of these mRNAs and enhances the fibrotic response. After searching for differential expression in TOF versus RV, we also observed 72 significantly differentially expressed miRNAs between TOF and LV. Many of them are significantly upregulated (in total 60) and only few are significantly downregulated (in total 12) in TOF compared to LV. Most of these miRNAs (88%) overlap with those significantly differentially expressed in TOF compared to RV, i.e. 55 of 60 upregulated miRNAs and 8 of 12 downregulated miRNAs. Differential expression analysis between right and left ventricle of healthy individuals revealed only three significant miRNAs namely the downregulated miR-223 and miR-142 as well as the upregulated miR-215 in RV compared to LV. miR-223 was also found most significantly downregulated in TOF versus LV. This miRNA regulates glucose transporter 4 (Glut4) protein expression and cardiomyocyte glucose metabolism[300]. miRNA-215 is significantly upregulated in RV versus LV and significantly downregulated in TOF compared to RV and can target WNK1[301]. It was shown that WNK1 ablation causes cardiovascular developmental defects[302]. Moreover, an essential role of endothelial WNK1 in the control of blood pressure and postnatal angiogenesis and cardiac growth was indicated by Xie *et al.*[302].

5.4.1 Novel MicroRNA Prediction

Approximately 30% of the mapped small RNA-seq reads could not be assigned to known miRNAs, other small non-coding RNAs, mRNAs or genomic repeats. Therefore, we searched for novel miRNAs over all samples using a fold- and scoring-based approach based on the miRDeep[190] package. Briefly, for novel miRNA prediction we used all sequences with a mapped read length of less or equal than 25 nucleotides (longer sequences are unlikely to represent mature miRNA sequences) as well as a sequence count of more than 25 (removing noise) which are not annotated to known miRNAs or other small non-coding RNAs resulting in ∼43 million reads representing ∼206,000 unique sequences.

To find novel miRNA candidates we used miRDeep considering clusters of reads that align along the reference genome, i.e. alignment pattern of the miRNA precursor sequence (mature miRNA sequence – loop sequence – star sequence) expected from miRNA processing. If such an alignment pattern was found, two potential precursor sequences (flanking regions of a mature miRNA sequence) were cut from the human

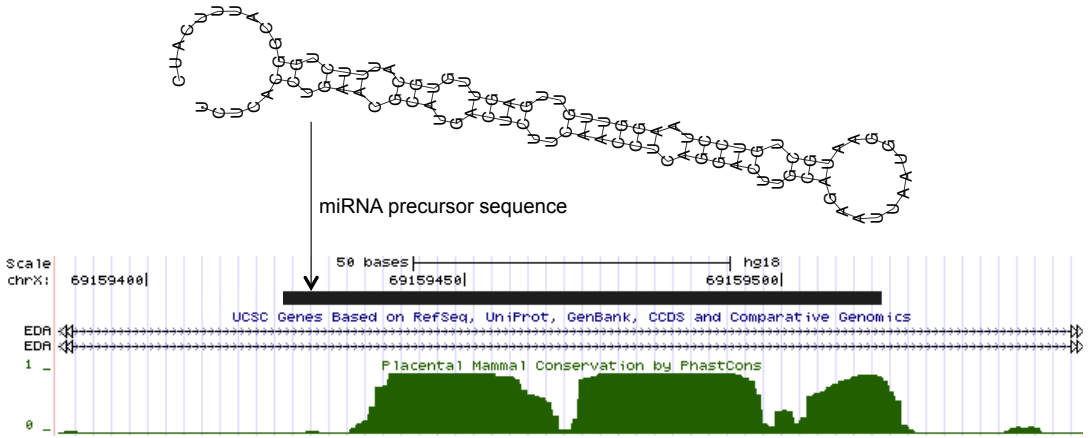

Figure 5.23: An example for a novel miRNA precursor sequence (located on chrX:69159420-69159516 based on NCBI v36.1, hg18). This miRNA precursor shows a good structure prediction by RNAfold (optimal secondary structure with a minimum free energy of -36.70 kcal/mol is given on top) and mammalian conservation based on PhastCons scores (conservation track from UCSC genome browser is given on the lower panel). In total, 230 reads (representing unique sequences with count > 25) could be mapped to this miRNA precursor sequence, i.e. 183 reads correspond to one distinct mature RNA and 47 reads correspond to one distinct mature* RNA. The mature/mature* duplex shows the typical 3' overhang. For the 5' arm, one read that has the correct position for a moRNA (directly adjacent to the mature* sequence) has been detected. For the 3' arm, no moRNA reads were found, but the conservation pattern indicates that there might be a conserved moRNA but not expressed in our samples (see conserved block 3' of the mature miRNA).

reference genome assuming that the mature sequence locates to the 5' arm or to the 3'arm of the RNA hairpin. Each potential miRNA precursor sequence was assessed after folding into a hairpin structure using the RNA folding algorithm from the Vienna[303] package. Furthermore, miRDeep searches for potential cleavage sites of Drosha and Dicer and uses the phylogenetic conservation as well as the filtering of other known small non-coding RNA species to improve the predictions. The stability of potential precursors sequences is tested using Randfold[304] v2.0. In summary, each potential miRNA precursor sequence was scored based on its read signature, secondary structure (e.g. multi-loops or a high minimum free energy decrease the score), cleavage, conservation and overlap to known small non-coding RNAs. In total, we found 100 novel miRNA candidates, of which 56 have annotated miRNA homologs in other species.

The novel miRNA candidates were further assessed by manual inspection. We searched

for well-formed secondary structures that contain a hairpin loop or only few bulges or internal loops. If both the mature and mature star sequence are detected, the processed ∼22 nt duplex should have a 3' overhang which is characteristic for Dicer processing. Additionally, a high percentage (≥75%) of reads should correspond to one or more distinct miRNA/miRNA* duplex showing a precise excision[305]. An exact 5' end processing is important especially since the nucleotides 2-7 comprise the seed sequence of the mature miRNA[306]. Potential miRNA loci were also checked for the expression of miRNA offset RNAs (moRNAs). These ∼20 nt RNAs are generated at a low level from sequences immediately adjacent to the mature miRNA and miRNA* (or even overlapping by few nucleotides). MoRNAs are especially found in evolutionary old miRNAs[307]. In addition, the conservation of the potential precursor sequences was evaluated using PhastCons conservation scores. Ideally, miRNAs show a high conservation for the arms and a lower conservation for the hairpin loop. Although conservation is widely considered as an important feature of miRNAs, it is not absolutely necessary for annotation. For example, Ambros *et al.* defined five expression and biogenesis criteria for annotation of miRNAs[308]. They stated that phylogenetic conservation is a stronger evidence for miRNA biogenesis than the prediction of a fold-back precursor. If only a predicted precursor but no conservation can be found, a miRNA can nevertheless be annotated if it is supported by strong expression data. In plants, Meyer *et al.* also showed that conservation is not necessary for annotation of miRNAs, although it provides especially strong evidence in favor of an annotation[305]. There is only one criterion that has to be fulfilled, i.e. that a 21 nt microRNA/microRNA* duplex is precisely excised from the stem of a single stranded, stem-loop precursor. As mentioned before, excision can be regarded as precise when more than 75% of observed small RNA abundance corresponds to one or more distinct miRNA/miRNA* duplexes.

Finally, we identified 33 potential novel precursor sequences (high confidence novel miRNAs) based on their frequency in human heart samples (e.g. there is one miRNA precursor sequence with over ∼50,000 mapped reads), predicted secondary structure and conservation (Supplementary Figure S12). An example for a high confidence novel miRNA precursor sequence is given in Figure 5.23. In addition, we examined their differential expression between TOF patients and right ventricle of healthy individuals. For the differential expression analysis the mapped reads of all samples were initially annotated by miRBase v14 and all novel miRNA annotations. Afterwards, we again performed differential expression analysis between TOF and RV and observed that almost all novel miRNAs were also upregulated in TOF. From our high confidence novel

miRNAs 16 are significantly upregulated in TOF versus RV (corrected p-value<0.05) and only two are significantly downregulated (Supplementary Figure S12). Because the miRBase repository is constantly updated in regular intervals, sequences will be updated or revised and new sequences will be added. Therefore, we checked whether our high confidence novel miRNAs are present in a newer version of miRBase and found that seven high confidence novel miRNAs are annotated in miRBase v15. This indicates that the performed novel miRNA prediction as well as our manual inspection revealed good results according to possible real novel miRNAs. However, the presence of predicted high confidence novel miRNAs should be further validated experimentally by e.g. qRT-PCR.

5.5 MicroRNA Target Prediction and Correlation Analysis

MicroRNA target prediction was performed for the 103 miRNAs which are significantly differentially expressed (corrected p-value<0.05) between TOF patients and right ventricle of healthy individuals. For prediction, we used the available predictions from three different tools including miRanda[269], PicTar[274] and TargetScan[257]. As all tools use quite different approaches (see Chapter 3.6.3) and sets of 3'UTR regions, the overlap between their target predictions is relatively small. In summary, we found 40,257, 15,206 and 18,418 predicted target transcripts for miRanda, PicTar and TargetScan, respectively. The miRanda algorithm uses more relaxed criteria (presumably resulting in a higher false positive rate but a lower false negative rate) as compared to PicTar or TargetScan, which accounts for the higher number of predictions for miRanda. In addition, more miRNAs were represented in miRanda over PicTar or TargetScan (80 miRNAs compared with 56 and 77 miRNAs, respectively). Searching for transcripts predicted by at least two of the three prediction tools resulted in 18,524 target transcripts for 78 miRNAs. For example, miR-215 was predicted by miRanda and TargetScan to target the WNK1 3'UTR. To further decrease the false positive rate we only looked at transcripts predicted by all three tools and found 8,875 transcripts representing 3,332 target genes of 54 miRNAs (10,071 miRNA-mRNA pairs). However, the number of predicted miRNA-mRNA pairs was still high. Therefore, we reduced this number further by searching for significantly differentially expressed mRNAs in TOF compared to RV and found 657 predicted miRNA-mRNA pairs representing 48 miRNAs and 216 mRNAs.

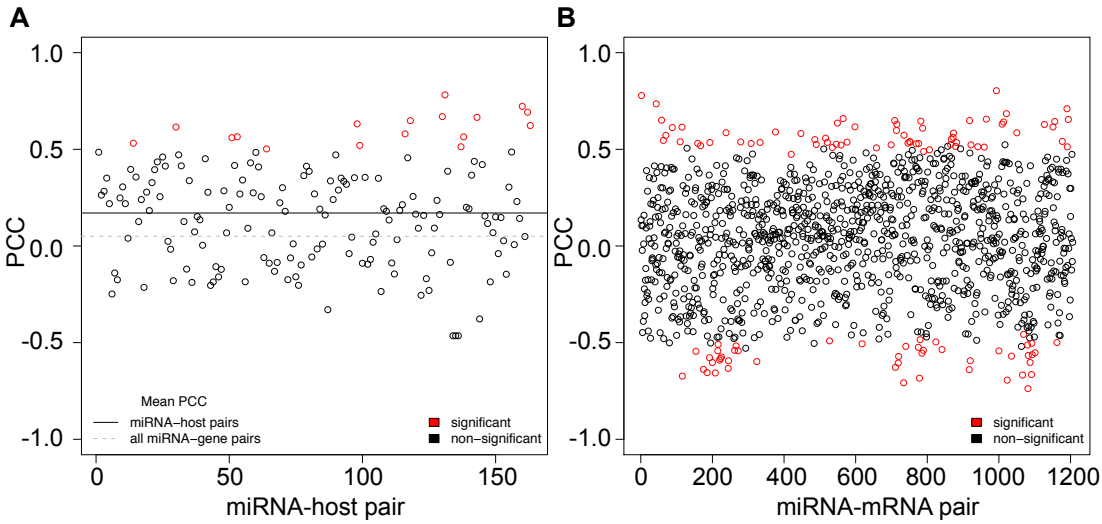

Figure 5.24: Correlation of miRNAs and (A) their host genes as well as (B) validated target genes. For Pearson correlation coefficient between expression levels in miRNA and their host or target genes all TOF patients with healthy individuals were used. Significance of miRNA-gene correlation ($p<0.05$) was assessed using random experiments in which expression values were shuffled across all individuals.

Of our 16 significant TOF genes (see Chapter 5.2, Figure 5.5A) only one gene is significantly differentially expressed (downregulated), namely the endothelin converting enzyme 2 (ECE2). ECE2 is a predicted target (by all three prediction tools used in this study) of the significantly upregulated miR-27b. This miRNA is differentially expressed from early stages of ventricular chamber formation [309] and promotes angiogenesis [310]. Moreover, it was shown that miR-27b targets NOTCH1 [311], a critical determinant of cardiac stem cell growth and differentiation [312]. NOTCH1 is also one of our potential TOF genes (see Chapter 5.2, Figure 5.5B) and downregulated in TOF compared to RV, although not statistically significant but with a fold change of 0.88.

In the past it has been shown that the expression level of many miRNAs can be both positively and negatively correlated with their target mRNAs [41]. For example, validated targets of miR-27b are CYP1B1 [313] and MEF2c [309]. CYP1B1 is significantly downregulated in our TOF patients versus RV. In contrast, MEF2c, an essential regulator of cardiac myogenesis and right ventricular development [314], is upregulated in TOF with a fold change of 1.34. For miR-19b there is a number of validated target genes showing differential expression between TOF and RV including SOCS-1 [315] (down-

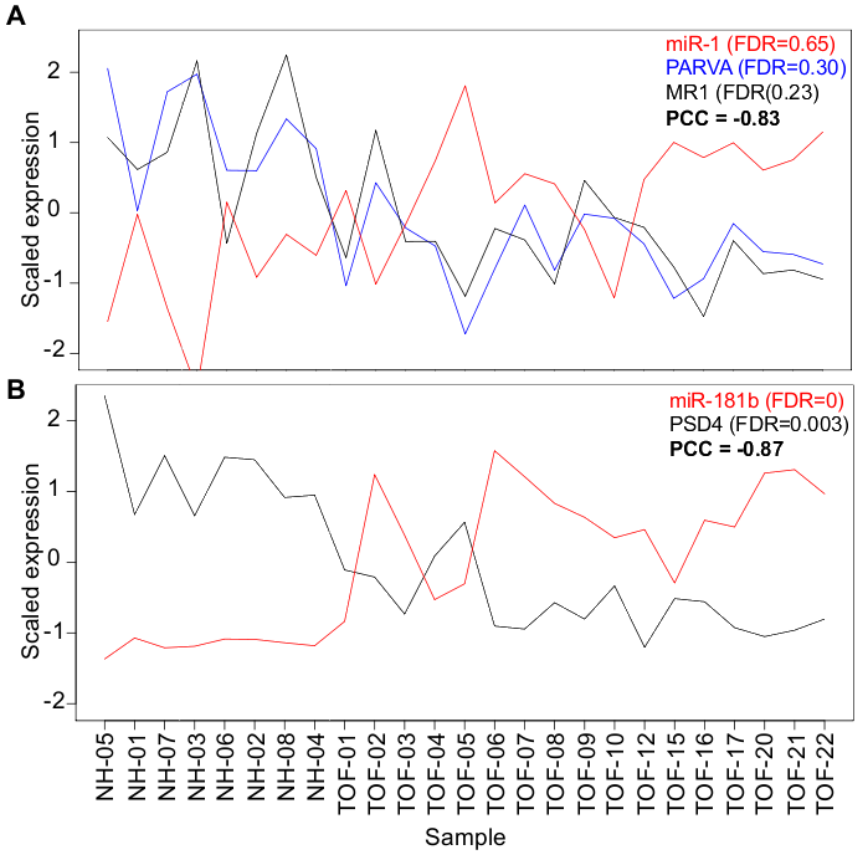

Figure 5.25: Examples for negative correlation based on the Pearson correlation coefficient (PCC) in scaled expression of miRNAs and genes over all individuals. The miRNA expression level is shown in red. Genes with a predicted binding site for the miRNA are shown in blue, otherwise in black. Differential expression between TOF patients and healthy individuals (normal heart, NH) is indicated by the false discovery rate (FDR). Significance was defined as FDR<0.05 after adjustment for multiple testing.

regulated with FC=0.71; prevents TNF-alpha-induced apoptosis in cardiac myocytes via ERK1/2 pathway activation[316]), PTEN[317] (upregulated with FC=1.1; involved in heart failure, myocardial hypertrophy and contractility[318]), VEGFA[319] (upregulated with FC=1.23; mutations are associated with congenital left ventricular outflow tract obstruction[320]) and ERalpha[321] (downregulated with FC=0.74; protective against the development of cardiac hypertrophy[322]). For further analysis, we computed Pearson correlation coefficients between expression levels of miRNAs and target or host genes using all TOF patients together with the healthy individuals. In general, the miRNAs

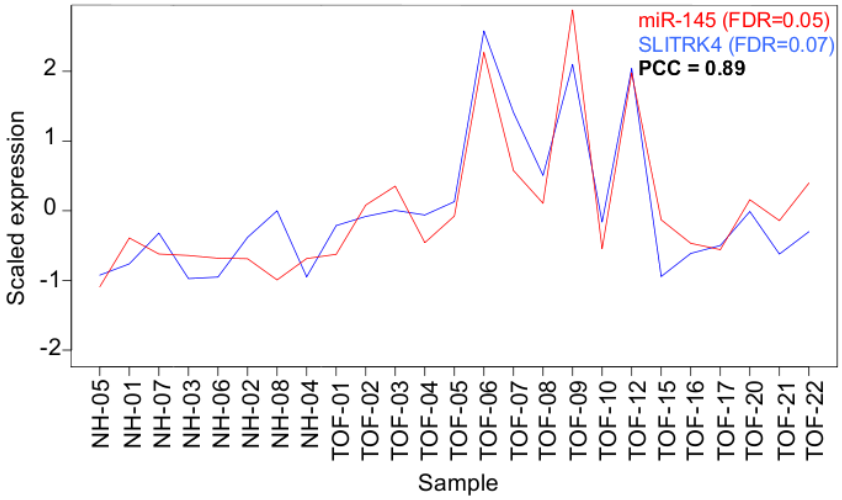

Figure 5.26: An example for a positive correlation in scaled expression of miR-145 (in red) and its target gene SLITRK4 (in blue) over all individuals. Differential expression between TOF patients and healthy individuals (normal heart, NH) is indicated by the false discovery rate (FDR). Significance was defined as FDR<0.05 after adjustment for multiple testing.

and their host genes (miRNA-host pair) show higher positive correlation in comparison to any miRNA-mRNA pairing (Figure 5.24A), although individual miRNA-host pairs can also show negative or no correlation. Looking at the correlation between miRNAs and validated human targets (based on miRecords[323] v3.0, TarBase[324] v5.0 and miRTarBase[325]) we found a broad variety of correlation ranging from significant positive to significant negative correlation (Figure 5.24B). Significance of miRNA-gene correlation ($p<0.05$) was assessed using random experiments in which expression values were shuffled across all individuals. Compared to any miRNA-mRNA pair, no clear shift to negative correlations was observed over all miRNAs. Looking at the (scaled) expression of individuals miRNAs and genes, we found pairings with both high negative as well as high positive correlation. miR-1 for example, which is highly expressed in skeletal muscle and heart, shows a very high negative correlation (Pearson correlation coefficient of -0.83) to PARVA and MR1 (Figure 5.25A). Both PARVA, which encodes a member of the parvin family of actin-binding proteins, and the myofibrillogenesis regulator MR-1 are also highly expressed in skeletal muscle and heart. But like miR-1, they are not significantly differentially expressed in TOF compared to RV due to different expression levels within one group. Only for PARVA a 3'UTR target site for miR-1 was predicted. A very high negative correlation (Pearson correlation coefficient of -0.87) was also ob-

miRNAs with negative correlation to their validated targets (neg. targets / all targets)	**miRNAs with negative correlation to their predicted targets** (neg. targets / all targets)
miR-1 (90/173)	miR-29a (169/269)
miR-27b (3/4)	miR-29b (206//268)
miR-29b (9/10)	miR-29c (212//269)
miR-29c (12/14)	miR-33a (14/41)
miR-204 (7/14)	miR-133b (96/124)
	miR-302b (169/195)
miRNAs with positive correlation to their validated targets (pos. targets / all targets)	**miRNAs with positive correlation to their predicted targets** (pos. targets / all targets)
miR-9 (6/7)	let-7b (137/227)
miR-21 (30/41)	let-7i (158/228)
	miR-9 (225/320)
	miR-27b (178/289)
	miR-92b (133/181)
	miR-101 (117/156)
	miR-130a (188/261)
	miR-152 (132/182)
	miR-181b (181/264)
	miR-203 (79/124)
	miR-208b (2/3)
	miR-218 (128/203)
	miR-221 (45/74)
	miR-222 (49/73)

Table 5.5: MicroRNAs with overall correlation in (scaled) expression to their validated (left column) and predicted (right column) target genes. For each miRNA the number of positively (pos.) or negatively (neg.) validated (based on miRecords, TarBase and/or miRTarBase) or predicted (by miRanda, PicTar and/or TargetScan) targets is given. Significant difference between targets and all genes (non-targets) is indicated by a t-test p-value, i.e. distribution over all pos. or neg. correlated miRNA-target pairs in comparison to all miRNA-non-target pairs. For the given miRNAs all p-values are smaller than 0.05.

served between miR-181b and PSD4 (Figure 5.25B). Moreover, the expression levels within one group (i.e. healthy or affected) are nearly equally distributed and both are differentially expressed (i.e. highly versus lowly expressed). Accordingly, this is associated with significant differential expression level between TOF and RV. Nevertheless, there is no prediction that the upregulated miR-181b targets the downregulated PSD4. An example for a very high positive correlation (Pearson correlation coefficient of 0.89)

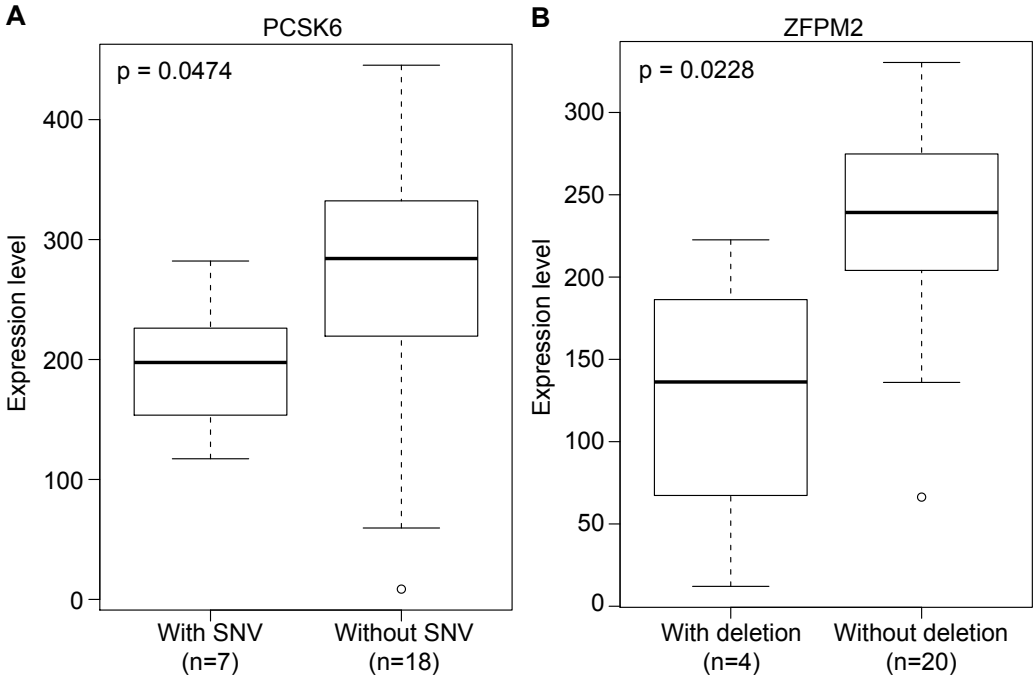

Figure 5.27: Local variations in predicted miRNA binding sites validated by Sanger sequencing in TOF patients as well as right ventricle of healthy individuals that potentially lead to a significant expression alteration between individuals with and without the local variation (Wilcoxon test with $p<0.05$). Expression levels are based on mRNA-seq data. (A) A single nucleotide variation (SNV) in PCSK6 (chr15:99662447, C>T, position based on NCBI v36.1, hg18) leads to loss of a binding site for miR-485 (miRanda prediction score without SNV = 142 and with SNV = 110). (B) A deletion in ZFPM2 (chr8:106885176, delGTTAT, position based on NCBI v36.1, hg18) leads to a novel binding site for miR-548j (miRanda prediction score without SNV = 125 and with SNV = 146).

in expression is miR-145 and its target gene SLITRK4 (Figure 5.26). They are highly positively correlated in their expression levels in both healthy and affected group. This miRNA-mRNA pair could be predicted by all three tools. Finally, we analyzed if certain miRNAs show an overall tendency to be negatively or positively correlated to their predicted as well as validated targets and found examples for both groups (Table 5.5).

After miRNA target prediction and correlation analysis we also searched for local variations in predicted miRNA binding sites that could lead to a significant gene expression alteration in patients showing a specific mutation compared to those not having the mutation in that gene (t-test with $p<0.05$). We examined all local variations found in

3'UTRs of genes over the 13 TOF patients with DNA, mRNA and miRNA sequencing data. For each mutation we searched for miRNAs with sufficient different miRanda prediction score in the reference sequence compared to the mutated sequence. The score was computed using a small window around the mutation (±20 bp) for both the reference and mutated sequence of the predicted target gene. A sufficient different miRanda prediction score was found if the score was greater or equal than 140 for one of the sequences and smaller for the other one and if the difference between both prediction scores (i.e. reference versus mutated sequence) was greater than 20. Finally, we found 85 local variations in predicted miRNA binding sites that lead to significant expression alterations, representing 99 miRNAs (all expressed in TOF, RV or in both) targeting 72 affected genes.

After manual assessment we selected two local variations (one SNV and one deletion) for validation by Sanger sequencing now in all 22 TOF patients (TOF-01 to TOF-22) and right ventricle of three healthy individuals (NH-02, NH-04 and NH-06) with available gene expression data (see Chapter 2.2.4, Table 2.2). First, a known SNV in PCSK6 was found that leads to loss of a binding site for miR-485 (miRanda prediction score without SNV = 142 and with SNV = 110). A significant expression alteration between individuals with and without this variation could be observed. Interestingly, instead of upregulation we found a significant downregulation of PCSK6 in individuals with this variation and the associated loss of the predicted binding site compared to the other individuals (average gene expression level in mRNA-seq of 170 mapped reads for individuals with the SNV and 290 mapped reads without the SNV; Figure 5.27A). This should be further analyzed and experimental validated by e.g. luciferase arrays. Second, a novel deletion in ZFPM2 leads to a novel binding site for miR-548j (miRanda prediction score without SNV = 125 and with SNV = 146). The predicted novel binding site leads to a significant downregulation of ZFPM2 in individuals with compared to those without such a deletion (average gene expression level in mRNA-seq of 136 mapped reads for individuals with the deletion and 256 mapped reads without the deletion; Figure 5.27B).

Finally, filtering all 85 local variations to be novel or with a MAF of less than or equal to 0.01 or present in OMIM resulted in just three local variations (one deletion and 2 SNVs) in three genes (SGCA, MTPN and ZFPM2). These novel variations were found in non-coding sequences related to predicted binding sites of five miRNAs (hsa-miR-548j, hsa-miR-15a, hsa-miR-16, hsa-miR-195 and hsa-miR-873).

Chapter 6

Discussion

The first part of this work represents a systematic *in vivo* analysis of three levels regulating cardiac mRNA profiles, namely regulation of gene transcription by epigenetic and genetic factors as well as post-transcriptional regulation by small non-coding RNAs. In detail, genome-wide binding of the key cardiac transcription factor Srf was analyzed in conjunction with functional consequences of RNAi mediated knockdown of Srf in cell culture, leading to new insights into its individual binding behavior and function. Further, Srf co-occurrence with histone 3 acetylation as well as the potential regulatory impact of miRNAs was studied.

In human and mouse approximately 2,000 transcription factors, more than 100 different modifications of histone residues and a large number of post-transcriptional regulators comprising over 1,000 miRNAs modulate the mRNA profile corresponding to 20,000-25,000 protein-coding genes. Major insights have been gained into the regulation of the transcriptional process by DNA-binding transcription factors and their modulators[7,326,327]. In addition, the role of histone modifications in establishing and maintaining the chromatin status and their function as protein interaction partners has been discovered[328–330]. More recently, the impact of miRNAs on mRNA profiles and their function as inhibitors of the translational process has emerged[169,331–333] as initial insights were obtained by focusing on each level independently. However, we lack data showing the interaction between these levels of regulation since the initial insights were obtained by focusing on each level independently. While it was long thought that transcription factors are the main driving force, results of this and other studies favor a comparable impact for all three regulatory levels with a high degree of interdependency leading to a fine-tuned balance of gene expression.

Investigating the influence of histone modifications as an epigenetic mechanism to modulate gene expression, we showed that the transcriptional activity of Srf in the mouse cardiomyocyte cell line HL-1 is highly depending on the co-occurrence of histone 3 acetylation. Using ChIP-chip it was previously shown by us that genes showing H3ac are less likely differentially expressed pointing to a buffering effect. To confirm this finding, we repeated the ChIP-chip experiments using the more sensitive ChIP-seq. Our data also revealed that the presence of H3ac tags had a buffering effect on the expression of Srf targets even after knockdown of this TF.

Interestingly, while both ChIP-chip and ChIP-seq approaches aim to measure the same enriched binding sites, a low overlap between the peak positions was found in Srf ChIP-chip compared to ChIP-seq data. This low overlap can have different reasons which have been further addressed in the community[288,334,335]. For example, a comparison of neuron-restrictive silencer factor (NRSF) peaks showed that only 22% of their ChIP-chip peaks overlapped with ChIP-seq peaks. However, the overlapping peaks had a much higher number of observed motifs than those that occurred only in ChIP-chip or ChIP-seq[193]. Summarizing this and other studies, the two methods 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 our observation, that the overlap of Srf peaks was much smaller than those of H3ac peaks, as the latter exhibit much stronger signals in the ChIP experiment. Besides the different experimental techniques, differences in the detected binding sites are also based on the algorithmic approaches used for peak calling. However, this explanation is unlikely as we observed a high overlap of 91% that was observed for H3ac. For ChIP-seq, no negative control sample was measured and thus, the background distribution was modeled from the ChIP sample itself using the negative binomial distribution. This distribution is more accurate than earlier approaches and it was shown that for a one-sample analysis, where only a ChIP sample is sequenced, reasonable FDR estimates can be provided[193]. Nevertheless, the repetitive analysis using ChIP-seq data revealed the same overall results as for the ChIP-chip data.

As mentioned before, Srf target gene activation was shown to be highly dependent on histone modifications. Histone modifying enzymes represent an important group of direct downstream targets of Srf as found in both ChIP-chip and ChIP-seq. For ex-

ample the histone demethylases containing a Jumonji domain such as Jmjd1c, Jmjd2b, Jmjd3, Jmjd4 and Jmjd5 were all found to be direct targets of Srf. A similar picture could be drawn for the relationship of miRNAs and Srf such that Argonaute proteins Eif2c2 (Ago2) and Eif2c (Ago3), which are direct Srf targets, play a key role for miRNA mediated-mRNA cleavage via the RISC complex[336]. In line with this, we found a panel of miRNAs deregulated in Srf knockdown, explaining three times more differentially expressed genes than Srf binding events alone could do. This further reflects the high degree of interdependency between the different levels.

The observed impact of H3ac on the activating potential of transcription factors like Srf underlines the beneficial effects seen for HDAC inhibitors for a variety of disease states[88]. Further, results from this study favor the view that modulation by histone modification as well as buffering by co-binding transcription factors 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 resulting in various cardiac malformations including atrial and ventricular septal defect [112,337]. The acetylation of histone 3 mediated via the histone acetyltransferase p300 provides an explanation for the observed high target gene expression of Srf. The correlation between Srf, p300 and H3ac was further investigated *in vivo* using ChIP-qPCR in a time-series during cardiac maturation in mouse[338]. In summary, a strong correlation between the occurrence of H3ac marks as well as Srf and p300 binding at potent regulatory regions of heart- and muscle-relevant genes was found. This points to a common regulatory mechanism which is triggered by Srf and resulted in H3ac that depends to a certain degree on the HAT p300[338].

In accordance with others, we observed that the overwhelming proportion of differentially expressed genes in our RNAi experiments were indirect targets of Srf. Computational studies suggest that up to 30% of all human genes are regulated by miRNAs, while each miRNA may control hundreds of target genes[339,340]. Our *in vivo* data highlight the global impact of miRNAs on expression profile alterations seen in transcription factor loss-of-function studies. Significantly differentially expressed miRNAs in Srf knockdown potentially explain up to 45% of the altered mRNA profile in our study. Over the last years a panel of miRNAs was discovered having a significant impact on the cardiac development and function. Differentially expressed miRNAs in Srf knockdown have been linked to vital processes such as arrhythmia (miR-1, miR-133),

apoptosis (miR-21, miR-195), contractility (miR-208), hypertrophy (miR-1, miR-21, miR-133, miR-195, miR-208) and fibrosis (miR-21)[341–347]. Furthermore, miR-1 promotes myogenesis by targeting HDAC4[95], a transcriptional repressor of muscle gene expression and thus represents an interface to histone acetylation.

To analyze miRNA-seq data, we developed MicroRazerS, a filter-based algorithm to map deep sequenced small RNAs to a reference genome. With the exponentially growing output of emerging deep sequencing platforms, fast and effective mapping of reads is a basic problem conserning a large community of researchers. MicroRazerS was compared with other short read mapping tools incorporating Mega BLAST[192] and the two possible best competitors Bowtie[176] and SOAP2[181]. We found MicroRazerS an order of magnitude faster or at least comparable in speed to the other short read mapping tools. In addition, it is more sensitive and easy to handle and adjust. Just recently it was shown that within six alignment tools tested, specifically devoted to miRNA detection, SHRiMP[182] and MicroRazerS showed the highest sensitivity[348]. Some useful options inherited from RazerS[174] are supplied like the option that counts uncalled nucleotides as automatic matches or the option that discards reads that map more than a designated number of times to the reference genome. Hence, MicroRazerS is an even more useful tool. Further, given the heterogenous nature of the small RNA types and the various output of sequencing platforms, it can be expected that mapping tools can to some degree work complementary thereby offering optimal solutions to distinct tasks.

In the second part of the work we studied Tetralogy of Fallot. TOF accounts for up to 10% of all CHD, which are the most common birth defects in human. Considering the background hypothesis of congenital heart disease, CHD are most likely caused by a panel of genetic variations with each effecting protein function or expression only modestly and manifest as disease only when combined with additional genetic, epigenetic or environmental alterations. To provide proof for this hypothesis we used latest next-generation sequencing techniques to discover genetic alterations in the cardiovascular exome and transcriptome of TOF cases, parents and controls. Further, we investigated genome-wide mRNA and miRNA levels in TOF cases and healthy unaffected individuals and combined gene expression profiles with miRNA target predictions.

Oligogenic disorders potentially represent a broad and significant number of diseases in general, which have been less accessible for conventional genetic studies. Therefore, only limited insight has been gained so far. Examples of known oligogenic disease are isolated gonadotropin-releasing hormone (GnRH) deficiency, Bardet-Biedl syndrome and

neural tube defects[349–351]. In this work, we show an oligogenic architecture of TOF with a mutation pattern characterized by a combination of common and rare alleles. We show that the observed mutation pattern in the TOF patients is very unlikely to occur in healthy controls. This provides a strong significant hint that the genes defined in the study are indeed reflecting the genetic background associated with the disease. Comparing the individual mutation pattern of each of our TOF patients to the control group revealed no healthy individual showing exactly the same combination of affected genes. This further underlines the importance of functionally interacting variations.

We identified SNVs and InDels in 16 genes that discriminate isolated TOF genotypes from those of healthy controls. These genes show a significantly higher mutation rate in TOF subjects compared to controls. On average, four TOF genes show deleterious mutations in an individual patient comprising novel and inherited mutations. This defines TOF as an oligogenic disorder. We found a characteristic mutation pattern in the TOF population. Out of 16 TOF genes two are affected in $\geq$50% of subjects, six genes in $\geq$20% and eight genes in $\geq$10% of subjects. We statistically assessed this pattern focusing on the ten most significant genes by a random permutation approach. We were unable to find any comparable mutation pattern in the control population, showing its statistical significance. Affected TOF genes harbor common and rare alleles showing a high dependency of functionally interacting yet individual mutations which lead grossly to the same phenotypic outcome during development. We postulate that these mutant alleles produce a genetic interaction network with abnormal properties that causes TOF.

To ascertain the complex genetic background of isolated TOF, we applied large-scale next-generation sequencing. The availability of control populations and the statistical assessment are key elements to extract indeed disease-relevant variations. Beside the significant TOF genes, we identified deleterious mutations in 221 additional genes of which 124 genes are mutated at a similar or higher frequency in controls compared to TOF patients and therefore are unlikely to be disease-causing. In total 97 genes were not assessable for statistical measures of which 30 genes were not targeted in controls and 67 harbor only InDels not studied in controls. It is likely that additional genes out of the set will turn out to be relevant to TOF and we described all genes affected in at least two TOF patients as potential TOF genes (in total 21).

The two most important modifier genes found were mitochondrial short-chain specific acyl-CoA dehydrogenase (ACADS, also known as SCAD) and titin, both well-known

genes in terms of mitochondropathy and cardiomyopathy[352–354]. The two observed and already known variations in ACADS show only a modest reduction in the enzymatic activity, but do not lead to clinically relevant SCAD deficiency on their own[290,355]. However, in combination with other genetic factors, the enzymatic activity could per se drop below the functionally needed critical threshold. In this study we provide evidence that this might be the situation in the affected patients, which show an altered PAS staining in their heart tissue, potentially suggesting a mitochondrial deficiency. TTN mutations on the other hand are associated with a panel of cardiomyopathies such as dilated or hypertrophic cardiomyopathy[356,357]. Like ACADS, all observed TTN mutations in our TOF patients occur in combination with other variations, suggesting that TTN as well as ACADS are important modifier genes. They occurred e.g. in combination with mutations of COL6A2. Variations in the COL6A1/COL6A2 cluster on chromosome 21 are associated with CHD in trisomy 21[358]. It was recently shown that overexpression of COL6A2 in combination with Down syndrome cell adhesion molecule (DSCAM) as a modifier gene can induce cardiac malformations in mouse[359].

A literature and database analysis as well as qRT-PCR and *in situ* hybridization of mouse hearts demonstrate the expression of TOF genes during heart development. This is essential to the hypothesis that TOF genes have a causative effect on abnormal cardiac development. Interestingly, in addition to the expression of TOF genes in embryonic development, we demonstrate a continued relevance during the postnatal period and adulthood. This is intriguing in respect to the differences that have been reported in the clinical outcome of TOF linked to the genetic background in syndromic cases[360].

Studying families with recurrent CHD, we show that respective mutations in TOF genes can be either novel or inherited, which explains incomplete penetrance in familiar cases[107]. Moreover, the genotype of healthy parents holds a significantly higher number of deleterious mutations compared to healthy non-CHD related controls. These data suggest that sequencing approaches can be integrated into genetic counseling for TOF to help determine risk profiles for individuals and families. The *a priori* identification of a risk profile in parents of offspring with TOF needs further exploration, particularly if this profile can be associated with other risk factors such as maternal diabetes or obesity[120,121]. Our data show that multiple genes provide the disease associated genetic background, and it frequently involves a disruption of signal transduction and metabolic pathways. For example, it has been shown previously that Nos3 (nitric oxide

synthase 3) genetically interacts with Tbx5 and plays a role in the development of atrial septal defects in Tbx5 knockout mice[361]. We show that NOS3 genetically interacts in TOF cases with the transcription factor TCF25 as well as plexin A2 (PLXNA2). NOS3 is regulated by many CHD risk factors including diabetes and provides an example how gene-environment interaction might interfere with human birth defects such as CHD. The impact of metabolic or environmental factors in combination with the parent genotype might permit the development of individual preventive strategies. Further studies are warranted to gather insights into key nodes and modulators of the genetic interaction network perturbed by TOF genes.

Digital gene expression information provided by RNA-seq can be used to validate local variations in coding regions and simultaneously assess the impact of such genetic variations on gene expression[285]. We gathered all mRNA-seq reads which mapped to found local variations and could validate ∼96% of them when using a minimal coverage of 10x in mRNA-seq, indicating high confidence local variations. Gene expression analysis revealed slightly more downregulated than upregulated genes in TOF patients compared to right ventricle of healthy individuals. Analyzing the gene expression similarity within the individual groups and between groups indicates a commonly changed expression profile in TOF patients. Further, we found TOF patients to be more similar in their gene expression to left ventricle of healthy individuals. This is in line with the results from Kaynak *et al.*, where the expression of several genes in right ventricular hypertrophy was similar to the expression in LV. A significant positive correlation was found, indicating that the genes dysregulated in right ventricular hypertrophy have a tendency to behave similarly in the disease state as in normal LV tissue[170].

Based on our gene expression profiles, we found the majority of the significant and potential TOF genes being expressed (RPKM>1), but only few of them significantly differentially expressed in TOF compared to normal heart. Genetic variations influencing gene expression may reside within the regulatory sequences, splice sites, secondary structure motifs and promoters or enhancers of the affected gene[362]. Especially sequence variations in promotor, enhancer and insulators (non-coding) regions should come into our focus for further studies as a putative cause of disturbed transcriptional regulation leading to congenital heart disease.

RNA-seq has been shown to be more sensitive compared to microarrays, both in terms of detection of lowly expressed and differentially expressed genes[144,211]. In fact, we found a high number of lowly expressed genes (RPKM≤1), which are also significantly

differentially expressed. However, these genes are more or less irrelevant, though they show randomly some significant differences in expression. The lowly expressed genes are expressed at less than one copy per cell on average and moreover, they are likely to correspond to 'leaky' expression, producing non-functional transcripts[293]. In many cases, differential regulation induces only small changes in expression levels, which probably serves to fine-tune expression[293]. However, many genes have a low and rather constant expression across tissues[363], indicating that our measured expression might be affected by subpopulations of cells. Using RNA-seq on single-cell level like in the study by Tang *et al.*[364], it will be possible to identify the core set of expressed genes in every individual cell.

Extracting biological insights from transcript-level RNA-seq analysis is challenging. Therefore, we also quantified isoforms using the POEM model comprising junction reads in the exonic read counts. However, the model can be extended to include junction reads in a more probabilistic way instead of adjusting just the corresponding exonic read counts[214,365]. Overall we found a high overlap between the gene-level and transcript-level results, although less significantly differentially lowly expressed genes (due to the lower read count) were observed based on the transcript-level analysis. This is in line with the fact that very lowly expressed transcripts in respect to their assigned read counts are discarded after POEM estimation, before they are tested for differential expression (low read count over all analyzed samples, i.e. they are very unlikely to be expressed).

Changes in the splicing machinery can be the cause of human diseases[366,367]. Analyzing alternative splicing in our RNA-seq data, we found novel splicing events in several sarcomeric genes. Among these genes it has been shown for e.g. TNNI1 and MYH7 that associated changes in mRNA splicing were significantly altered in patients with ischemic cardiomyopathy, dilated cardiomyopathy and aortic stenosis[295]. Moreover, mutations in MYH7 are associated with familial hypertrophic cardiomyopathy[368]. The candidate novel splice sites in both genes could be validated by RT-PCR as well as for PDLIM3, which is involved in cytoskeletal assembly and colocalizes with alpha-actinin-2 (ACTN2) at the Z lines of skeletal muscle[369]. PDLIM3 regulates SRF activity and isoform ratios play a role in muscle cell differentiation[370]. However, we evaluated the overall impact of differential splicing as a potential disease-causing mechanism and found only few significantly differentially transcripts related to differential splicing events. Moreover, no deleterious mutations was found on a splicing factor.

Post-transcriptional regulation of gene expression by miRNAs plays an important role in multiple cellular pathways and diseases. Deep sequencing of miRNAs in TOF patients revealed mostly upregulated miRNAs compared to normal heart. In heart failure the majority of miRNAs was also found to be upregulated and the expression profile was found to be similar to fetal hearts[371]. Several heart- and muscle-relevant miRNAs could be identified as significantly differentially expressed like in other studies investigating heart diseases[344,347,371,372]. An important feature of miRNAs is the ability to regulate the produced protein level of a multitude of mRNAs. Several computational tools have been developed for predicting miRNA targets[256]. Unfortunately, all prediction tools use different approaches and sets of 3'UTRs. Consequently, the amount of overlapping miRNA-mRNA predictions is often low, although each of the tools can identify a large number of potential miRNA targets. Using the overlap of three commonly used prediction tools for only significantly differentially expressed miRNAs and mRNAs in TOF patients compared to normal heart we found a reasonable number of miRNA-mRNA pairs. This number was further reduced if we only retain pairs with negatively correlated expression levels. However, it has been shown that the expression level on many miRNAs can be both positively and negatively correlated with their individual target genes[41]. Looking at the correlation between miRNAs and validated targets we found both significant positive and negative correlation. Compared to any miRNA-mRNA pair, no clear tendency to negative correlation was observed over all miRNAs. Looking at the expression of individual miRNAs and target genes, we again found predicted pairings with both high negative and high positive correlation, although the positively correlated pairs were slightly predominate.

Finally, we searched for local variations in predicted miRNA binding sites that lead to a significant gene expression alteration in the affected TOF patients compared to those without the mutation. We found an already known single nucleotide variation in PCSK6 that potentially leads to the loss of a predicted binding site for miR-485. Interestingly, we found a significant downregulation of PCSK6 in the TOF patients with this variation, which should be further analyzed. PCSK6 is a serine endoprotease that can cleave precursor proteins and it has been shown that its knockout in mouse leads to severe cardiac defects like persistent truncus arteriosus, ventricle septum defect and abnormal heart looping[373]. We also found and validated a novel deletion in ZFPM2 that leads to a predicted novel binding site for miR-548j. TOF patients with this deletion showed a significant downregulation of ZFPM2 in our mRNA-seq data. ZFPM2 (or FOG-2) is a zink finger protein that regulates activity of GATA transcrip-

tion factors. Moreover, ZFPM2 is essential for heart morphogenesis. ZFPM2 knockout embryos die at midgestation with a cardiac defect characterized by a thin ventricular myocardium, common atrioventricular canal and TOF malformation[374]. The relevance of both predicted miRNA binding sites needs to be further analyzed.

Compared to microarrays, expression values obtained from mRNA-seq correlate better with protein levels. However, the expression levels correlate not perfectly due to post-transcriptional regulation[375]. In this work we searched for genetic alterations in coding regions of over ∼1,000 heart- and muscle-relevant genes and miRNAs and combined genome-wide data from mRNA and small RNA sequencing to identify potential TOF genes and miRNAs as their post-transcriptional modifiers. In the future we need to look not only at the RNA level but also at the protein level, because the relationship between RNA levels and protein levels varies[376,377]. It has been shown that there is a correlation between mRNA levels and protein concentrations[378,379] and moreover, we could try to model the contribution of general sequence features[380]. However, as these predictions are so far only partially reliable for a meaningful statement we have to measure protein levels.

In this thesis next-generation sequencing technologies have been extensively used to discover different players of gene expression. Prospectively however, NGS technologies will be replaced more and more by single-molecule sequencing approaches (third-generation sequencing)[381,382], that will further increase throughput with even longer reads (promising more than 1 kbp) than any other technology at present. Longer reads will improve the data quality including read mapping, base calling (polymorphism detection) and *de novo* assembly. With higher dimensional data we may evolve an even more complete understanding of living systems and complex phenotypes like congenital heart disease. However, we provide proof for the long-standing hypothesis that CHD are in part caused by an underlying oligogenic background and report an advance for analyzing oligo- or multigenic disorders using the recent NGS technologies. Studying TOF, we used a small cohort of patients and families with recurrent CHD. To further substantiate our findings, future studies should incorporate a larger number of patients and families. Nevertheless, we are convinced that our analysis strategy and bioinformatics approach provides valuable insights into the causes of CHD and can be applied to other oligo- or multigenic disorders in general.

Bibliography

[1] J. Schlesinger*, M. Schueler*, M. Grunert*, J. J. Fischer*, Q. Zhang, T. Krueger, M. Lange, M. Tönjes, I. Dunkel, and S. R. Sperling, "The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs.," *PLoS genetics*, vol. 7, p. e1001313, Feb. 2011. *Equal contributions. i, 21, 25, 59, 60, 70

[2] A.-K. Emde*, M. Grunert*, D. Weese, K. Reinert, and S. R. Sperling, "MicroRazerS: rapid alignment of small RNA reads.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 123–124, Jan. 2010. *Equal contributions. i, 31

[3] M. Grunert*, C. Dorn*, M. Schueler*, I. Dunkel, J. Schlesinger, S. Mebus, K. Bellmann, V. Alexi-Meskishvili, S. Klaassen, K. Wassilew, B. Timmermann, R. Hetzer, F. Berger, and S. R. Sperling, "Dissecting Congenital Heart Disease - The Oligogenic Basis of Isolated Tetralogy of Fallot.," *Under review.* *Equal contributions. i

[4] F. Crick, "Central dogma of molecular biology.," *Nature*, vol. 227, pp. 561–563, Aug. 1970. 1

[5] W. Gilbert, "Why genes in pieces?," *Nature*, vol. 271, p. 501, Feb. 1978. 1

[6] B. Lemon and R. Tjian, "Orchestrated response: a symphony of transcription factors for gene control.," *Genes & development*, vol. 14, pp. 2551–2569, Oct. 2000. 2

[7] 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. Genetics*, vol. 10, pp. 252–263, Apr. 2009. 2, 113

[8] F. J. Asturias, "RNA polymerase II structure, and organization of the preinitiation complex.," *Current opinion in structural biology*, vol. 14, pp. 121–129, Apr. 2004. 2

[9] M. Levine and R. Tjian, "Transcription regulation and animal diversity.," *Nature*, vol. 424, pp. 147–151, July 2003. 2

[10] S. Ogbourne and T. M. Antalis, "Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes.," *The Biochemical journal*, vol. 331 (Pt 1), pp. 1–14, Apr. 1998. 2

[11] J.-J. M. Riethoven, "Regulatory regions in DNA: promoters, enhancers, silencers, and insulators.," *Methods in molecular biology (Clifton, N.J.)*, vol. 674, pp. 33–42, 2010. 2

[12] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 A resolution.," *Nature*, vol. 389, pp. 251–260, Sept. 1997. 2

[13] L. Ho and G. R. Crabtree, "Chromatin remodelling during development.," *Nature*, vol. 463, pp. 474–484, Jan. 2010. 2

[14] M. M. Suzuki and A. Bird, "DNA methylation landscapes: provocative insights from epigenomics.," *Nature reviews. Genetics*, vol. 9, pp. 465–476, June 2008. 2

[15] Y. Wang, J. Wysocka, J. R. Perlin, L. Leonelli, C. D. Allis, and S. A. Coonrod, "Linking covalent histone modifications to epigenetics: the rigidity and plasticity of the marks.," *Cold Spring Harbor symposia on quantitative biology*, vol. 69, pp. 161–169, 2004. 2

[16] S. L. Berger, "The complex language of chromatin regulation during transcription.," *Nature*, vol. 447, pp. 407–412, May 2007. 2

[17] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications.," *Nature*, vol. 403, pp. 41–45, Jan. 2000. 2

[18] 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*, vol. 51, pp. 786–794, May 1964. 3

[19] S. R. Bhaumik, E. Smith, and A. Shilatifard, "Covalent modifications of histones during development and disease pathogenesis.," *Nature structural & molecular biology*, vol. 14, pp. 1008–1016, Nov. 2007. 3

[20] N. A. Faustino and T. A. Cooper, "Pre-mRNA splicing and human disease.," *Genes & development*, vol. 17, pp. 419–437, Feb. 2003. 3

[21] H. Richard, M. H. Schulz, M. Sultan, A. Nürnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, S. A. Haas, and M.-L. Yaspo, "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments.," *Nucleic acids research*, vol. 38, p. e112, June 2010. 3, 18, 40

[22] B. J. Blencowe, "Alternative splicing: new insights from global analyses.," *Cell*, vol. 126, pp. 37–47, July 2006. 3

[23] R. Lee and R. Feinbaum, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," *Cell*, 1993. 3

[24] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun, "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.," *Nature*, vol. 403, pp. 901–906, Feb. 2000. 3

[25] M. Lagos-Quintana, R. Rauhut, and W. Lendeckel, "Identification of novel genes coding for small expressed RNAs," *Science (New York, N.Y.)*, 2001. 3

[26] H. B. Houbaviy, M. F. Murray, and P. A. Sharp, "Embryonic Stem Cell-Specific MicroRNAs," *Developmental Cell*, vol. 5, pp. 351–358, Aug. 2003.

[27] N. C. Lau, "An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans," *Science (New York, N.Y.)*, vol. 294, pp. 858–862, Oct. 2001. 3

[28] "An extensive class of small RNAs in Caenorhabditis elegans," 2001. 3

[29] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel, "The microRNAs of Caenorhabditis elegans.," *Genes & development*, vol. 17, pp. 991–1008, Apr. 2003.

[30] Z. Mourelatos, J. Dostie, S. Paushkin, A. Sharma, B. Charroux, L. Abel, J. Rappsilber, M. Mann, and G. Dreyfuss, "miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs.," *Genes & development*, vol. 16, pp. 720–728, Mar. 2002.

[31] B. J. Reinhart, E. G. Weinstein, M. W. Rhoades, B. Bartel, and D. P. Bartel, "MicroRNAs in plants.," *Genes & development*, vol. 16, pp. 1616–1626, July 2002.

[32] M. C. Vella and F. J. Slack, "C. elegans microRNAs.," *WormBook : the online review of C. elegans biology*, pp. 1–9, 2005. 3

[33] H. Kawaji and Y. Hayashizaki, "Exploration of small RNAs.," *PLoS genetics*, vol. 4, p. e22, Jan. 2008. 3

[34] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley, "Identification of mammalian microRNA host genes and transcription units.," *Genome research*, vol. 14, pp. 1902–1910, Oct. 2004. 3

[35] S. Baskerville, "Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes," *Rna*, vol. 11, pp. 241–247, Jan. 2005.

[36] Y.-K. Kim and V. N. Kim, "Processing of intronic microRNAs," *The EMBO journal*, vol. 26, pp. 775–783, Jan. 2007.

[37] X. Cai, C. H. Hagedorn, and B. R. Cullen, "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs.," *Rna*, vol. 10, pp. 1957–1966, Dec. 2004. 3

[38] F. Fazi and C. Nervi, "MicroRNA: basic mechanisms and transcriptional regulatory networks for cell fate determination.," *Cardiovascular research*, vol. 79, pp. 553–561, Sept. 2008. 4

[39] L. Guo and Z. Lu, "The Fate of miRNA* Strand through Evolutionary Analysis: Implication for Degradation As Merely Carrier Strand or Potential Regulatory Molecule?," *PloS one*, vol. 5, p. e11387, June 2010. 4

[40] M. Inui, G. Martello, and S. Piccolo, "MicroRNA control of signal transduction.," *Nature reviews. Molecular cell biology*, vol. 11, pp. 252–263, Apr. 2010. 5, 53

[41] J. Nunez-Iglesias, C.-C. Liu, T. E. Morgan, C. E. Finch, and X. J. Zhou, "Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation.," *PloS one*, vol. 5, no. 2, p. e8898, 2010. 53, 107, 121

[42] A. Arvey, E. Larsson, C. Sander, C. S. Leslie, and D. S. Marks, "Target mRNA abundance dilutes microRNA and siRNA activity.," *Molecular systems biology*, vol. 6, p. 363, Apr. 2010. 5, 53

[43] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data.," *Nucleic acids research*, vol. 39, pp. D152–7, Jan. 2011. 5, 42, 64, 66, 98

[44] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel, "Most mammalian mRNAs are conserved targets of microRNAs.," *Genome research*, vol. 19, pp. 92–105, Jan. 2009. 5, 54

[45] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen, "Principles of MicroRNA–Target Recognition," *PLoS biology*, vol. 3, p. e85, Feb. 2005. 5

[46] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson, "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.," *Nature*, vol. 433, pp. 769–773, Feb. 2005. 5, 53

[47] F. Sanger, "DNA sequencing with chain-terminating inhibitors," Jan. 1977. 5

[48] L. Bonetta, "Genome sequencing in the fast lane," *Nature methods*, vol. 3, pp. 141–147, Feb. 2006. 5

[49] S. Schuster, "Next-generation sequencing transforms today's biology," *Nature*, 2007. 5

[50] J. Shendure, "Next-generation DNA sequencing," *Nat Biotechnol*, 2008. 5, 6, 7

[51] O. Harismendy, P. Ng, and R. Strausberg, "Evaluation of next generation sequencing platforms for population targeted sequencing studies," *Genome . . .*, 2009. 5, 9, 57

[52] M. L. Metzker, "Sequencing technologies — the next generation," *Nature reviews. Genetics*, vol. 11, pp. 31–46, Dec. 2009. 6

[53] 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 research*, vol. 36, p. e105, Sept. 2008.

[54] Illumina, Inc, "Official website for Illumina, Inc.," Oct. 2011. 6

[55] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch, "Accuracy and quality of massively parallel DNA pyrosequencing.," *Genome biology*, vol. 8, no. 7, p. R143, 2007. 6

[56] K. V. Voelkerding, S. A. Dames, and J. D. Durtschi, "Next-generation sequencing: from basic research to diagnostics.," *Clinical chemistry*, vol. 55, pp. 641–658, Apr. 2009.

[57] . L. Science, "Official website for 454 Life Science (Roche)," Oct. 2011. 6, 55

[58] A. Biosystems, "Official website for Applied Biosystems," Oct. 2011. 6

[59] D. Pushkarev, N. F. Neff, and S. R. Quake, "Single-molecule sequencing of an individual human genome.," *Nat Biotechnol*, vol. 27, pp. 847–850, Sept. 2009. 6

[60] Helicos BioSciences Corporation, "Official website for Helicos BioSciences Corporation," Oct. 2011. 6

[61] L. Kruglyak and D. A. Nickerson, "Variation is the spice of life.," *Nature genetics*, vol. 27, pp. 234–236, Mar. 2001. 8

[62] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome.," *Nature reviews. Genetics*, vol. 7, pp. 85–97, Feb. 2006. 8

[63] Illumina, "Empowering GWAS for a new era of discovery.," Feb. 2012. 8

[64] Y. Xue, Q. Wang, Q. Long, B. L. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, Asan, D. G. MacArthur, M. A. Quail, N. P. Carter, H. Yang, and C. Tyler-Smith, "Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree.," *Current biology : CB*, vol. 19, pp. 1453–1457, Sept. 2009. 8

[65] K. K. Linask and J. W. Lash, "Early heart development: dynamics of endocardial cell sorting suggests a common origin with cardiomyocytes.," *Developmental dynamics : an official publication of the American Association of Anatomists*, vol. 196, pp. 62–69, Jan. 1993. 9

[66] S. Martin-Puig, Z. Wang, and K. R. Chien, "Lives of a Heart Cell: Tracing the Origins of Cardiac Progenitors," *Cell Stem Cell*, vol. 2, pp. 320–331, Apr. 2008. 9

[67] S. M. Wu, K. R. Chien, and C. Mummery, "Origins and fates of cardiovascular progenitor cells.," *Cell*, vol. 132, pp. 537–543, Feb. 2008. 9

[68] K. R. Chien, I. J. Domian, and K. K. Parker, "Cardiogenesis and the complex biology of regenerative cardiovascular medicine.," *Science (New York, N.Y.)*, vol. 322, pp. 1494–1497, Dec. 2008. 9

[69] 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*, vol. 271, pp. H2183–9, Nov. 1996. 9

[70] B. G. Bruneau, "The developmental genetics of congenital heart disease," *Nature*, vol. 451, pp. 943–948, Feb. 2008. 9, 11

[71] M. Ruiz, "Tetralogy of Fallot," June 2006. 10, 12

[72] E. N. Olson, "Gene regulatory networks in the evolution and development of the heart.," *Science (New York, N.Y.)*, vol. 313, pp. 1922–1927, Sept. 2006. 9, 11

[73] J. K. Takeuchi, M. Ohgi, K. Koshiba-Takeuchi, H. Shiratori, I. Sakaki, K. Ogura, Y. Saijoh, and T. Ogura, "Tbx5 specifies the left/right ventricles and ventricular septum position during cardiogenesis.," *Development (Cambridge, England)*, vol. 130, pp. 5953–5964, Dec. 2003. 10

[74] Y.-S. Dai, P. Cserjesi, B. E. Markham, and J. D. Molkentin, "The transcription factors GATA4 and dHAND physically interact to synergistically activate cardiac gene expression through a p300-dependent mechanism.," *The Journal of biological chemistry*, vol. 277, pp. 24390–24398, July 2002. 10

[75] Y. Lee, T. Shioi, H. Kasahara, S. M. Jobe, R. J. Wiese, B. E. Markham, and S. Izumo, "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*, vol. 18, pp. 3120–3129, June 1998. 10

[76] S. Morin, F. Charron, L. Robitaille, and M. Nemer, "GATA-dependent recruitment of MEF2 proteins to target promoters.," *The EMBO journal*, vol. 19, pp. 2046–2055, May 2000. 10

[77] V. Garg, I. S. Kathiriya, R. Barnes, M. K. Schluterman, I. N. King, C. A. Butler, C. R. Rothrock, R. S. Eapen, K. Hirayama-Yamada, K. Joo, R. Matsuoka, J. C. Cohen, and D. Srivastava, "GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5.," *Nature*, vol. 424, pp. 443–447, July 2003. 10

[78] N. S. Belaguli, J. L. Sepulveda, V. Nigam, F. Charron, M. Nemer, and R. J. Schwartz, "Cardiac tissue enriched factors serum response factor and GATA-4 are mutual coregulators.," *Molecular and cellular biology*, vol. 20, pp. 7550–7558, Oct. 2000. 10

[79] J. M. Miano, X. Long, and K. Fujiwara, "Serum response factor: master regulator of the actin cytoskeleton and contractile apparatus.," *American journal of physiology. Cell physiology*, vol. 292, pp. C70–81, Jan. 2007. 10

[80] 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*, vol. 272, pp. 18222–18231, July 1997. 10

[81] J. M. Miano, "Serum response factor: toggling between disparate programs of gene expression.," *Journal of molecular and cellular cardiology*, vol. 35, pp. 577–593, June 2003. 10

[82] D. Wang, P. S. Chang, Z. Wang, L. Sutherland, J. A. Richardson, E. Small, P. A. Krieg, and E. N. Olson, "Activation of cardiac gene expression by myocardin, a transcriptional cofactor for serum response factor.," *Cell*, vol. 105, pp. 851–862, June 2001. 11

[83] G. Posern and R. Treisman, "Actin' together: serum response factor, its cofactors and the link to signal transduction.," *Trends in cell biology*, vol. 16, pp. 588–596, Nov. 2006. 10

[84] J. L. Sepulveda, S. Vlahopoulos, D. Iyer, N. Belaguli, and R. J. Schwartz, "Combinatorial expression of GATA4, Nkx2-5, and serum response factor directs early cardiac gene activity.," *The Journal of biological chemistry*, vol. 277, pp. 25775–25782, July 2002. 10

[85] Z. Wang, D.-Z. Wang, G. C. T. Pipes, and E. N. Olson, "Myocardin is a master regulator of smooth muscle gene expression.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 7129–7134, June 2003. 11

[86] B. G. Bruneau, "Transcriptional regulation of vertebrate cardiac morphogenesis.," *Circulation research*, vol. 90, pp. 509–519, Mar. 2002. 11

[87] J. K. Takeuchi, M. Mileikovskaia, K. Koshiba-Takeuchi, A. B. Heidt, A. D. Mori, E. P. Arruda, M. Gertsenstein, R. Georges, L. Davidson, R. Mo, C.-C. Hui, R. M. Henkelman, M. Nemer, B. L. Black, A. Nagy, and B. G. Bruneau, "Tbx20 dose-dependently regulates transcription factor networks required for mouse heart and motoneuron development.," *Development (Cambridge, England)*, vol. 132, pp. 2463–2474, May 2005. 11

[88] 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. Genetics*, vol. 10, pp. 32–42, Jan. 2009. 11, 115

[89] J. Lu, T. A. McKinsey, C. L. Zhang, and E. N. Olson, "Regulation of skeletal myogenesis by association of the MEF2 transcription factor with class II histone deacetylases.," *Mol Cell . . .*, vol. 6, pp. 233–244, Aug. 2000. 11

[90] D. Cao, Z. Wang, C.-L. Zhang, J. Oh, W. Xing, S. Li, J. A. Richardson, D.-Z. Wang, and E. N. Olson, "Modulation of smooth muscle gene expression by association of histone acetyltransferases and deacetylases with myocardin.," *Molecular and cellular biology*, vol. 25, pp. 364–376, Jan. 2005. 11, 60

[91] T. Kouzarides, "Chromatin modifications and their function.," *Cell*, vol. 128, pp. 693–705, Feb. 2007. 11

[92] 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*, vol. 278, pp. 20047–20058, May 2003. 11

[93] K. R. Cordes, N. T. Sheehy, M. P. White, E. C. Berry, S. U. Morton, A. N. Muth, T.-H. Lee, J. M. Miano, K. N. Ivey, and D. Srivastava, "miR-145 and miR-143 regulate smooth muscle cell fate and plasticity.," *Nature*, vol. 460, pp. 705–710, Aug. 2009. 11

[94] C. Kwon, Z. Han, E. N. Olson, and D. Srivastava, "MicroRNA1 influences cardiac differentiation in Drosophila and regulates Notch signaling.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 18986–18991, Dec. 2005. 11

[95] J.-F. Chen, E. M. Mandel, J. M. Thomson, Q. Wu, T. E. Callis, S. M. Hammond, F. L. Conlon, and D.-Z. Wang, "The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation.," *Nature genetics*, vol. 38, pp. 228–233, Feb. 2006. 11, 23, 63, 116

[96] Z. Niu, A. Li, S. X. Zhang, and R. J. Schwartz, "Serum response factor micromanaging cardiogenesis.," *Current opinion in cell biology*, vol. 19, pp. 618–627, Dec. 2007.

[97] Y. Zhao, J. F. Ransom, A. Li, V. Vedantham, M. von Drehle, A. N. Muth, T. Tsuchihashi, M. T. McManus, R. J. Schwartz, and D. Srivastava, "Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2.," *Cell*, vol. 129, pp. 303–317, Apr. 2007. 11

[98] J. I. E. Hoffman and S. Kaplan, "The incidence of congenital heart disease.," *Journal of the American College of Cardiology*, vol. 39, pp. 1890–1900, June 2002. 11

[99] C. L. Webb, K. J. Jenkins, P. P. Karpawich, A. F. Bolger, R. M. Donner, H. D. Allen, R. J. Barst, and Congenital Cardiac Defects Committee of the American Heart Association Section on Cardiovascular Disease in the Young, "Collaborative care for adults with congenital heart disease.," *Circulation*, vol. 105, pp. 2318–2323, May 2002. 11

[100] J. I. Hoffman, "Incidence of congenital heart disease: II. Prenatal incidence.," *Pediatric cardiology*, vol. 16, pp. 155–165, June 1995. 12

[101] H. Yagi, Y. Furutani, H. Hamada, T. Sasaki, S. Asakawa, S. Minoshima, F. Ichida, K. Joo, M. Kimura, S.-i. Imamura, N. Kamatani, K. Momma, A. Takao, M. Nakazawa, N. Shimizu, and R. Matsuoka, "Role of TBX1 in human del22q11.2 syndrome.," *Lancet*, vol. 362, pp. 1366–1373, Oct. 2003. 12

[102] L. A. Jerome and V. E. Papaioannou, "DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1.," *Nature genetics*, vol. 27, pp. 286–291, Mar. 2001. 12

[103] Q. Y. Li, R. A. Newbury-Ecob, J. A. Terrett, D. I. Wilson, A. R. Curtis, C. H. Yi, T. Gebuhr, P. J. Bullen, S. C. Robson, T. Strachan, D. Bonnet, S. Lyonnet, I. D. Young, J. A. Raeburn, A. J. Buckler, D. J. Law, and J. D. Brook, "Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family.," *Nature genetics*, vol. 15, pp. 21–29, Jan. 1997. 12

[104] C. T. Basson, D. R. Bachinsky, R. C. Lin, T. Levi, J. A. Elkins, J. Soults, D. Grayzel, E. Kroumpouzou, T. A. Traill, J. Leblanc-Straceski, B. Renault, R. Kucherlapati, J. G. Seidman, and C. E. Seidman, "Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome.," *Nature genetics*, vol. 15, pp. 30–35, Jan. 1997. 12

[105] H. Matsson, J. Eason, C. S. Bookwalter, J. Klar, P. Gustavsson, J. Sunnegårdh, H. Enell, A. Jonzon, M. Vikkula, I. Gutierrez, J. Granados-Riveron, M. Pope, F. Bu'Lock, J. Cox, T. E. Robinson, F. Song, D. J. Brook, S. Marston, K. M. Trybus, and N. Dahl, "Alpha-cardiac actin mutations produce atrial septal defects.," *Human molecular genetics*, vol. 17, pp. 256–265, Jan. 2008. 13

[106] A. Mégarbané, N. Salem, E. Stephan, R. Ashoush, D. Lenoir, V. Delague, R. Kassab, J. Loiselet, and P. Bouvagnet, "X-linked transposition of the great arteries and incomplete penetrance among males with a nonsense mutation in ZIC3.," *European journal of human genetics : EJHG*, vol. 8, pp. 704–708, Sept. 2000. 13

[107] Z. A. Eldadah, A. Hamosh, N. J. Biery, R. A. Montgomery, M. Duke, R. Elkins, and H. C. Dietz, "Familial Tetralogy of Fallot caused by mutation in the jagged1 gene.," *Human molecular genetics*, vol. 10, pp. 163–169, Jan. 2001. 13, 118

[108] D. Srivastava, T. Thomas, Q. Lin, M. L. Kirby, D. Brown, and E. N. Olson, "Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND.," *Nature genetics*, vol. 16, pp. 154–160, June 1997. 13

[109] L. Zhu, R. Vranckx, P. Van Kien, and A. Lalande, "Mutations in myosin heavy chain 11 cause a syndrome associating thoracic aortic aneurysm/aortic dissection and patent ductus arteriosus," *Nature genetics*, 2006. 13

[110] J. J. Schott, D. W. Benson, C. T. Basson, W. Pease, G. M. Silberbach, J. P. Moak, B. J. Maron, C. E. Seidman, and J. G. Seidman, "Congenital heart disease caused by mutations in the transcription factor NKX2-5.," *Science (New York, N.Y.)*, vol. 281, pp. 108–111, July 1998. 13

[111] V. Garg, A. N. Muth, J. F. Ransom, M. K. Schluterman, R. Barnes, I. N. King, P. D. Grossfeld, and D. Srivastava, "Mutations in NOTCH1 cause aortic valve disease.," *Nature*, vol. 437, pp. 270–274, Sept. 2005. 13

[112] S. Sperling, C. H. Grimm, I. Dunkel, S. Mebus, H.-P. Sperling, A. Ebner, R. Galli, H. Lehrach, C. Fusch, F. Berger, and S. Hammer, "Identification and functional analysis of CITED2 mutations in patients with congenital heart defects.," *Human mutation*, vol. 26, pp. 575–582, Dec. 2005. 13, 115

[113] G. Nemer, F. Fadlalah, J. Usta, M. Nemer, G. Dbaibo, M. Obeid, and F. Bitar, "A novel mutation in the GATA4 gene in patients with Tetralogy of Fallot.," *Human mutation*, vol. 27, pp. 293–294, Mar. 2006. 13

[114] D. W. Benson, G. M. Silberbach, A. Kavanaugh-McHugh, C. Cottrill, Y. Zhang, S. Riggs, O. Smalls, M. C. Johnson, M. S. Watson, J. G. Seidman, C. E. Seidman, J. Plowden, and J. D. Kugler, "Mutations in the cardiac transcription factor NKX2.5 affect diverse cardiac developmental pathways.," *The Journal of clinical investigation*, vol. 104, pp. 1567–1573, Dec. 1999. 13

[115] S. A. Mohamed, Z. Aherrahrou, H. Liptau, A. W. Erasmi, C. Hagemann, S. Wrobel, K. Borzym, H. Schunkert, H. H. Sievers, and J. Erdmann, "Novel missense mutations (p.T596M and p.P1797H) in NOTCH1 in patients with bicuspid aortic valve.," *Biochemical and biophysical research communications*, vol. 345, pp. 1460–1465, July 2006. 13

[116] W. Gong, "Mutation analysis of TBX1 in non-deleted patients with features of DGS/VCFS or isolated cardiovascular defects," *Journal of Medical Genetics*, vol. 38, pp. 45e–45, Dec. 2001. 13

[117] E. P. Kirk, M. Sunde, M. W. Costa, S. A. Rankin, O. Wolstein, M. L. Castro, T. L. Butler, C. Hyun, G. Guo, R. Otway, J. P. Mackay, L. B. Waddell, A. D. Cole, C. Hayward, A. Keogh, P. Macdonald, L. Griffiths, D. Fatkin, G. F. Sholler, A. M. Zorn, M. P. Feneley, D. S. Winlaw, and R. P. Harvey, "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*, vol. 81, pp. 280–291, Aug. 2007. 13

[118] J. Bentham and S. Bhattacharya, "Genetic mechanisms controlling cardiovascular development.," *Annals of the New York Academy of Sciences*, vol. 1123, pp. 10–19, Mar. 2008. 13

[119] D. B. McElhinney, E. Geiger, J. Blinder, D. Woodrow Benson, and E. Goldmuntz, "NKX2.5 mutations in patients with congenital heart disease," *Journal of the American College of Cardiology*, vol. 42, pp. 1650–1655, Nov. 2003. 13

[120] C. A. Loffredo, P. D. Wilson, and C. Ferencz, "Maternal diabetes: an independent risk factor for major cardiovascular malformations with increased mortality of affected infants.," *Teratology*, vol. 64, pp. 98–106, Aug. 2001. 13, 118

[121] M. Watkins, S. Rasmussen, and M. Honein, "Maternal obesity and risk for birth defects," *Pediatrics*, 2003. 118

[122] V. Lopez and C. Keen, "Prenatal zinc deficiency: influence on heart morphology and distribution of key heart proteins in a rat model," *Biological trace element research*, 2008.

[123] P. Kastner, N. Messaddeq, M. Mark, O. Wendling, J. M. Grondona, S. Ward, N. Ghyselinck, and P. Chambon, "Vitamin A deficiency and mutations of RXRalpha, RXRbeta and RARalpha lead to early differentiation of embryonic ventricular cardiomyocytes.," *Development (Cambridge, England)*, vol. 124, pp. 4749–4758, Dec. 1997.

[124] A. STARREVELDZIMMERMAN, W. VANDERKOLK, J. ELSHOVE, and H. MEINARDI, "Teratogenicity of antiepileptic drugs," *Clinical Neurology and Neurosurgery*, vol. 77, pp. 81–95, Dec. 1974.

[125] E. Zimmerman, "Substance abuse in pregnancy: teratogenesis.," *Pediatric annals*, 1991.

[126] 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*, vol. 153, pp. 529–536, Mar. 2001. 13

[127] J. J. Nora, "Multifactorial inheritance hypothesis for the etiology of congenital heart diseases. The genetic-environmental interaction.," *Circulation*, vol. 38, pp. 604–617, Sept. 1968. 13

[128] J. J. Nora and A. H. Nora, "Recurrence risks in children having one parent with a congenital heart disease.," *Circulation*, vol. 53, pp. 701–702, Apr. 1976. 13

[129] A. Visel, E. M. Rubin, and L. A. Pennacchio, "Genomic views of distant-acting enhancers," *Nature*, vol. 461, pp. 199–205, Sept. 2009. 16

[130] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions.," *Science (New York, N.Y.)*, vol. 316, pp. 1497–1502, June 2007. 15

[131] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.," *Nature methods*, vol. 4, pp. 651–657, Aug. 2007. 15

[132] M. K. D. B. J. B. I. E. L. G. G. P. A. W. B. T.-K. K. R. P. K. W. L. E. M. A. O. A. P. C. R. X. X. A. M. M. W. R. J. C. N. E. S. L. B. E. B. Tarjei S Mikkelsen, "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells," *Nature*, vol. 448, p. 553, Aug. 2007. 15, 32, 33

[133] A. Barski, S. Cuddapah, K. Cui, T. Roh, and D. Schones, "High-resolution profiling of histone methylations in the human genome," *Cell*, 2007. 15

[134] P. J. Park, "ChIP-seq: advantages and challenges of a maturing technology.," *Nature reviews. Genetics*, vol. 10, pp. 669–680, Oct. 2009. 16, 17, 18, 33, 34, 35, 61, 62

[135] B. Ren, "Genome-Wide Location and Function of DNA Binding Proteins," *Science (New York, N.Y.)*, vol. 290, pp. 2306–2309, Dec. 2000. 17

[136] S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras, "Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.," *Cell*, vol. 116, pp. 499–509, Feb. 2004. 17

[137] T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren, "A high-resolution map of active promoters in the human genome," *Nature*, vol. 436, pp. 876–880, June 2005. 17

[138] 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*, vol. 7, p. 133, 2006. 17

[139] C. Schönbach, "From masking repeats to identifying functional repeats in the mouse transcriptome.," *Briefings in bioinformatics*, vol. 5, pp. 107–117, June 2004. 17

[140] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein, "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.," *Nat Biotechnol*, vol. 27, pp. 66–75, Jan. 2009. 17, 33

[141] N. Whiteford, N. Haslam, G. Weber, A. Prügel-Bennett, J. W. Essex, P. L. Roach, M. Bradley, and C. Neylon, "An analysis of the feasibility of short read sequencing.," *Nucleic acids research*, vol. 33, no. 19, p. e171, 2005. 17

[142] L. W. Hillier, G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, J. I. Glasscock, M. Hickenbotham, W. Huang, V. J. Magrini, R. J. Richt, S. N. Sander, D. A. Stewart, M. Stromberg, E. F. Tsung, T. Wylie, T. Schedl, R. K. Wilson, and E. R. Mardis, "Whole-genome sequencing and variant discovery in C. elegans," *Nature methods*, vol. 5, pp. 183–188, Jan. 2008. 18

[143] M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner, "A large genome center's improvements to the Illumina sequencing system.," *Nature methods*, vol. 5, pp. 1005–1010, Dec. 2008. 18

[144] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews. Genetics*, vol. 10, pp. 57–63, Jan. 2009. 18, 19, 119

[145] R. Morin, M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. Pugh, H. McDonald, R. Varhol, S. Jones, and M. Marra, "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing.," *BioTechniques*, vol. 45, pp. 81–94, July 2008. 37, 38

[146] C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A. M. Chinnaiyan, "Transcriptome sequencing to detect gene fusions in cancer.," *Nature*, vol. 458, pp. 97–101, Mar. 2009.

[147] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq.," *Nature methods*, vol. 5, pp. 621–628, July 2008. 37, 38, 44

[148] N. Cloonan, A. R. R. Forrest, G. Kolle, B. B. A. Gardiner, G. J. Faulkner, M. K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A. C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M. Manning, K. J. McKernan, and S. M. Grimmond, "Stem cell transcriptome profiling via massive-scale mRNA sequencing.," *Nature methods*, vol. 5, pp. 613–619, July 2008. 18, 19, 44

[149] M. Boguski and C. Tolstoshev, "Gene discovery in dbEST," *Science (New York, N.Y.)*, 1994. 18

[150] M. Hu and K. Polyak, "Serial analysis of gene expression," *Nature Protocols*, vol. 1, pp. 1743–1760, Nov. 2006. 18

[151] M. Harbers and P. Carninci, "Tag-based approaches for transcriptome research and genome annotation," *Nature methods*, vol. 2, pp. 495–502, July 2005. 18

[152] R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, Y. Hayashizaki, and P. Carninci, "CAGE: cap analysis of gene expression," *Nature methods*, vol. 3, pp. 211–222, Mar. 2006. 18

[153] S. Brenner, M. Johnson, J. Bridgham, and G. Golda, "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays," *Nature*, 2000. 18

[154] L. Mamanova, A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner, "Target-enrichment strategies for next-generation sequencing," *Nature methods*, vol. 7, pp. 111–118, Feb. 2010. 19, 20

[155] M. J. Clark, R. Chen, H. Y. K. Lam, K. J. Karczewski, R. Chen, G. Euskirchen, A. J. Butte, and M. Snyder, "Performance comparison of exome DNA sequencing technologies," *Nat Biotechnol*, vol. 29, pp. 908–914, Sept. 2011. 19

[156] Roche NimbleGen, Inc., "NimbleGen Sequence Capture technology," Oct. 2011. 20

[157] R. Tewhey, J. B. Warner, M. Nakano, B. Libby, M. Medkova, P. H. David, S. K. Kotsopoulos, M. L. Samuels, J. B. Hutchison, J. W. Larson, E. J. Topol, M. P. Weiner, O. Harismendy, J. Olson, D. R. Link, and K. A. Frazer, "Microdroplet-based PCR enrichment for large-scale targeted sequencing.," *Nat Biotechnol*, vol. 27, pp. 1025–1031, Nov. 2009. 19

[158] G. J. Porreca, K. Zhang, J. B. Li, B. Xie, D. Austin, S. L. Vassallo, E. M. LeProust, B. J. Peck, C. J. Emig, F. Dahl, Y. Gao, G. M. Church, and J. Shendure, "Multiplex amplification of large sets of human exons.," *Nature methods*, vol. 4, pp. 931–936, Nov. 2007.

[159] H. Johansson, M. Isaksson, E. F. Sörqvist, F. Roos, J. Stenberg, T. Sjöblom, J. Botling, P. Micke, K. Edlund, S. Fredriksson, H. G. Kultima, O. Ericsson, and M. Nilsson, "Targeted resequencing of candidate genes using selector probes.," *Nucleic acids research*, vol. 39, p. e8, Jan. 2011.

[160] A. Gnirke, A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum, "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.," *Nat Biotechnol*, vol. 27, pp. 182–189, Feb. 2009. 19

[161] T. J. Albert, M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler, X. Song, T. A. Richmond, C. M. Middle, M. J. Rodesch, C. J. Packard, G. M. Weinstock, and R. A. Gibbs, "Direct selection of human genomic loci by microarray hybridization.," *Nature methods*, vol. 4, pp. 903–905, Nov. 2007. 19

[162] D. T. Okou, K. M. Steinberg, C. Middle, D. J. Cutler, T. J. Albert, and M. E. Zwick, "Microarray-based genomic selection for high-throughput resequencing.," *Nature methods*, vol. 4, pp. 907–909, Nov. 2007.

[163] E. Hodges, Z. Xuan, V. Balija, M. Kramer, M. N. Molla, S. W. Smith, C. M. Middle, M. J. Rodesch, T. J. Albert, G. J. Hannon, and W. R. McCombie, "Genome-wide in situ exon capture for selective resequencing.," *Nature genetics*, vol. 39, pp. 1522–1527, Dec. 2007. 19

[164] B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. Kulbokas, T. R. Gingeras, S. L. Schreiber, and E. S. Lander, "Genomic maps and comparative analysis of histone modifications in human and mouse.," *Cell*, vol. 120, pp. 169–181, Jan. 2005. 21

[165] D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, and R. A. Young, "Genome-wide map of nucleosome acetylation and methylation in yeast.," *Cell*, vol. 122, pp. 517–527, Aug. 2005.

[166] B. E. Bernstein, E. L. Humphrey, R. L. Erlich, R. Schneider, P. Bouman, J. S. Liu, T. Kouzarides, and S. L. Schreiber, "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*, vol. 99, pp. 8695–8700, June 2002.

[167] D. Schübeler, D. M. MacAlpine, D. Scalzo, C. Wirbelauer, C. Kooperberg, F. van Leeuwen, D. E. Gottschling, L. P. O'Neill, B. M. Turner, J. Delrow, S. P. Bell, and M. Groudine, "The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote.," *Genes & development*, vol. 18, pp. 1263–1271, June 2004. 21

[168] J. J. Fischer, J. Toedling, T. Krueger, M. Schueler, W. Huber, and S. Sperling, "Combinatorial effects of four histone modifications in transcription and differentiation.," *Genomics*, vol. 91, pp. 41–51, Jan. 2008. 21

[169] E. van Rooij, N. Liu, and E. N. Olson, "MicroRNAs flex their muscles.," *Trends in genetics : TIG*, vol. 24, pp. 159–166, Apr. 2008. 23, 63, 113

[170] B. Kaynak, A. von Heydebreck, S. Mebus, D. Seelow, S. Hennig, J. Vogel, H.-P. Sperling, R. Pregla, V. Alexi-Meskishvili, R. Hetzer, P. E. Lange, M. Vingron, H. Lehrach, and S. Sperling, "Genome-wide array analysis of normal and malformed human hearts.," *Circulation*, vol. 107, pp. 2467–2474, May 2003. 25, 119

[171] S. Hammer, M. Toenjes, M. Lange, J. J. Fischer, I. Dunkel, S. Mebus, C. H. Grimm, R. Hetzer, F. Berger, and S. Sperling, "Characterization of TBX20 in human hearts and its regulation by TFAP2.," *Journal of cellular biochemistry*, vol. 104, pp. 1022–1033, June 2008.

[172] M. Toenjes, M. Schueler, S. Hammer, U. J. Pape, J. J. Fischer, F. Berger, M. Vingron, and S. Sperling, "Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes.," *Molecular bioSystems*, vol. 4, pp. 589–598, June 2008. 25

[173] C. Trapnell and S. L. Salzberg, "How to map billions of short reads onto genomes.," *Nat Biotechnol*, vol. 27, pp. 455–457, May 2009. 28

[174] D. Weese, A.-K. Emde, T. Rausch, A. Döring, and K. Reinert, "RazerS–fast read mapping with sensitivity control.," *Genome research*, vol. 19, pp. 1646–1654, Sept. 2009. 28, 29, 30, 61, 88, 116

[175] M. Ruffalo, T. Laframboise, and M. Koyutürk, "Comparative analysis of algorithms for next-generation sequencing read alignment.," *Bioinformatics (Oxford, England)*, vol. 27, pp. 2790–2796, Oct. 2011. 29

[176] 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*, vol. 10, no. 3, p. R25, 2009. 29, 31, 68, 116

[177] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1754–1760, July 2009. 29, 72

[178] A. J. Cox, "ELAND." 29

[179] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores.," *Genome research*, vol. 18, pp. 1851–1858, Nov. 2008. 29

[180] Novocraft, "Novoalign," Feb. 2012. 29

[181] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang, "SOAP2: an improved ultrafast tool for short read alignment.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1966–1967, Aug. 2009. 29, 31, 68, 116

[182] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno, "SHRiMP: accurate mapping of short color-space reads.," *PLoS computational biology*, vol. 5, p. e1000386, May 2009. 29, 116

[183] H. Lin, Z. Zhang, M. Q. Zhang, B. Ma, and M. Li, "ZOOM! Zillions of oligos mapped," *Bioinformatics (Oxford, England)*, vol. 24, pp. 2431–2437, Oct. 2008. 29

[184] O. Owolabi and D. R. McGregor, "Fast approximate string matching," *Software: Practice and Experience*, vol. 18, pp. 387–393, Apr. 1988. 29

[185] P. Jokinen, "Two algorithms for approxmate string matching in static texts," *Mathematical Foundations of Computer Science ...*, 1991. 29

[186] J. Karkkainen, "One-gapped q-gram filters for Levenshtein distance," *Combinatorial pattern matching*, 2002. 29

[187] S. Burkhardt, A. Crauser, P. Ferragina, H.-P. Lenhof, E. Rivals, and M. Vingron, "Proceedings of the third annual international conference on Computational molecular biology - RECOMB '99," in *the third annual international conference*, (New York, New York, USA), pp. 77–83, ACM Press, 1999. 29

[188] K. R. Rasmussen, J. Stoye, and E. W. Myers, "Efficient q-gram filters for finding all epsilon-matches over a given length.," *Journal of computational biology : a journal of computational molecular cell biology*, vol. 13, pp. 296–308, Mar. 2006. 29

[189] A. Döring, D. Weese, T. Rausch, and K. Reinert, "SeqAn an efficient, generic C++ library for sequence analysis.," *BMC bioinformatics*, vol. 9, p. 11, 2008. 29

[190] M. R. Friedländer, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky, "Discovering microRNAs from deep sequencing data using miRDeep.," *Nat Biotechnol*, vol. 26, pp. 407–415, Apr. 2008. 30, 103

[191] R. D. Morin, M. D. O'Connor, M. Griffith, F. Kuchenbauer, A. Delaney, A.-L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst, C. J. Eaves, and M. A. Marra, "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.," *Genome research*, vol. 18, pp. 610–621, Apr. 2008. 30

[192] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences.," *Journal of computational biology : a journal of computational molecular cell biology*, vol. 7, pp. 203–214, Jan. 2000. 30, 68, 116

[193] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong, "An integrated software system for analyzing ChIP-chip and ChIP-seq data.," *Nat Biotechnol*, vol. 26, pp. 1293–1300, Nov. 2008. 32, 33, 114

[194] 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 research*, vol. 36, pp. 5221–5231, Sept. 2008. 33

[195] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng, "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1952–1958, Aug. 2009. 33

[196] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-Seq (MACS).," *Genome biology*, vol. 9, no. 9, p. R137, 2008. 33

[197] G. Tuteja, P. White, J. Schug, and K. H. Kaestner, "Extracting transcription factor targets from ChIP-Seq data.," *Nucleic acids research*, vol. 37, p. e113, Sept. 2009. 32

[198] A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. M. Jones, "FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology.," *Bioinformatics (Oxford, England)*, vol. 24, pp. 1729–1730, Aug. 2008. 32, 33

[199] V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC: transcriptional regulation, from patterns to profiles.," *Nucleic acids research*, vol. 31, pp. 374–378, Jan. 2003. 35

[200] D. Vlieghe, A. Sandelin, P. J. De Bleser, K. Vleminckx, W. W. Wasserman, F. van Roy, and B. Lenhard, "A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.," *Nucleic acids research*, vol. 34, pp. D95–7, Jan. 2006. 35

[201] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences.," *Nucleic acids research*, vol. 18, pp. 6097–6100, Oct. 1990. 35

[202] M. L. Bulyk, "Computational prediction of transcription-factor binding site locations.," *Genome biology*, vol. 5, no. 1, p. 201, 2003. 36

[203] A. E. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, "MATCH: A tool for searching transcription factor binding sites in DNA sequences.," *Nucleic acids research*, vol. 31, pp. 3576–3579, July 2003. 35

[204] S. Rahmann, T. Müller, and M. Vingron, "On the power of profiles for transcription factor binding site detection.," *Statistical applications in genetics and molecular biology*, vol. 2, p. Article7, 2003. 35

[205] J.-V. Turatsinze, M. Thomas-Chollier, M. Defrance, and J. van Helden, "Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules.," *Nature Protocols*, vol. 3, no. 10, pp. 1578–1588, 2008. 35

[206] H. G. Roider, A. Kanhere, T. Manke, and M. Vingron, "Predicting transcription factor affinities to DNA from a biophysical model.," *Bioinformatics (Oxford, England)*, vol. 23, pp. 134–141, Jan. 2007. 35

[207] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. J. Searle, "Ensembl 2012.," *Nucleic acids research*, Nov. 2011. 37, 56, 88

[208] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye, "Database resources of the National Center for Biotechnology Information.," *Nucleic acids research*, vol. 39, pp. W528–32, Dec. 2011. 37, 56

[209] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, "RNA-Seq gene expression estimation with read mapping uncertainty.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 493–500, Feb. 2010. 37, 38, 39

[210] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by RNA sequencing.," *Science (New York, N.Y.)*, vol. 320, pp. 1344–1349, June 2008. 37

[211] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.," *Genome research*, vol. 18, pp. 1509–1517, Sept. 2008. 37, 44, 119

[212] T. Hashimoto, M. J. L. de Hoon, S. M. Grimmond, C. O. Daub, Y. Hayashizaki, and G. J. Faulkner, "Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2613–2614, Oct. 2009. 38, 88

[213] G. J. Faulkner, A. R. R. Forrest, A. M. Chalk, K. Schroder, Y. Hayashizaki, P. Carninci, D. A. Hume, and S. M. Grimmond, "A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE.," *Genomics*, vol. 91, pp. 281–288, Mar. 2008. 38

[214] H. Jiang and W. H. Wong, "Statistical inferences for isoform expression in RNA-Seq.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1026–1032, Apr. 2009. 38, 40, 120

[215] A. Dempster, "Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm," Feb. 1977. 40

[216] A. Cameron, "Regression analysis of count data," 1998. 40

[217] S. Audic and J. M. Claverie, "The significance of digital gene expression profiles.," *Genome research*, vol. 7, pp. 986–995, Oct. 1997. 40

[218] T. Beissbarth, L. Hyde, G. K. Smyth, C. Job, W.-M. Boon, S.-S. Tan, H. S. Scott, and T. P. Speed, "Statistical modeling of sequencing errors in SAGE libraries.," *Bioinformatics (Oxford, England)*, vol. 20 Suppl 1, pp. i31–9, Aug. 2004. 40

[219] P. Fujita, B. Rhead, and A. Zweig, "The UCSC Genome Browser database: update 2011," *Nucleic acids ...*, 2011. 42, 56, 99

[220] H. J. Peltier and G. J. Latham, "Normalization of microRNA expression levels in quantitative RT-PCR assays: identification of suitable reference RNA targets in normal and cancerous human solid tissues.," *Rna*, vol. 14, pp. 844–852, May 2008. 42

[221] T. Horie, K. Ono, H. Nishi, Y. Iwanaga, K. Nagao, M. Kinoshita, Y. Kuwabara, R. Takanabe, K. Hasegawa, T. Kita, and T. Kimura, "MicroRNA-133 regulates the expression of GLUT4 by targeting KLF15 and is involved in metabolic control in cardiac myocytes.," *Biochemical and biophysical research communications*, vol. 389, pp. 315–320, Nov. 2009.

[222] H. K. Kim, Y. S. Lee, U. Sivaprasad, A. Malhotra, and A. Dutta, "Muscle-specific microRNA miR-206 promotes muscle differentiation.," *The Journal of cell biology*, vol. 174, pp. 677–687, Aug. 2006. 42

[223] S. Ma and Y. Dai, "Principal component analysis based methods in bioinformatics studies.," *Briefings in bioinformatics*, vol. 12, pp. 714–722, Nov. 2011. 43

[224] J. Ramsay, "Maximum likelihood estimation in multidimensional scaling," *Psychometrika*, 1977. 43

[225] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data.," *Genome biology*, vol. 11, no. 3, p. R25, 2010. 44, 45, 46

[226] S. Linsen, E. De Wit, G. Janssens, and S. Heater, "Limitations and possibilities of small RNA digital gene expression profiling," *Nature methods*, 2009. 44, 45

[227] J. Bullard, E. Purdom, K. Hansen, and S. Durinck, *Statistical inference in mRNA-Seq: exploratory data analysis and differential expression*. UC Berkeley Division of . . . , 2009. 44

[228] M. D. Robinson and G. K. Smyth, "Small-sample estimation of negative binomial dispersion, with applications to SAGE data," *Biostatistics*, vol. 9, pp. 321–332, July 2007. 44

[229] P. A. C. t Hoen, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. A. M. Vossen, R. X. de Menezes, J. M. Boer, G.-J. B. van Ommen, and J. T. den Dunnen, "Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms.," *Nucleic acids research*, vol. 36, p. e141, Dec. 2008. 44

[230] R. Z. N. Vêncio, H. Brentani, D. F. C. Patrão, and C. A. B. Pereira, "Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE).," *BMC bioinformatics*, vol. 5, p. 119, Aug. 2004. 44

[231] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.," *Science (New York, N.Y.)*, vol. 321, pp. 956–960, Aug. 2008. 45

[232] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.," *BMC bioinformatics*, vol. 11, p. 94, 2010. 45

[233] J. T. Marques, K. Kim, P.-H. Wu, T. M. Alleyne, N. Jafari, and R. W. Carthew, "Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila," *Nature structural & molecular biology*, vol. 17, pp. 24–30, Dec. 2009. 45

[234] A. Git, H. Dvinge, M. Salmon-Divon, and M. Osborne, "Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression," *Rna*, 2010. 45

[235] S. U. Meyer, M. W. Pfaffl, and S. E. Ulbrich, "Normalization strategies for microRNA profiling experiments: a 'normal' way to a hidden layer of complexity?," *Biotechnology letters*, vol. 32, pp. 1777–1788, Dec. 2010. 46

[236] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 139–140, Jan. 2010. 46, 47, 88

[237] M. D. Robinson and G. K. Smyth, "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics (Oxford, England)*, vol. 23, pp. 2881–2887, Nov. 2007. 46, 47, 48

[238] K. A. Baggerly, L. Deng, J. S. Morris, and C. M. Aldaz, "Differential expression in SAGE: accounting for normal between-library variation.," *Bioinformatics (Oxford, England)*, vol. 19, pp. 1477–1483, Aug. 2003. 46

[239] K. A. Baggerly, L. Deng, J. S. Morris, and C. M. Aldaz, "Overdispersed logistic regression for SAGE: modelling multiple groups and covariates.," *BMC bioinformatics*, vol. 5, p. 144, Oct. 2004. 46

[240] J. Lu, J. K. Tomfohr, and T. B. Kepler, "Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach.," *BMC bioinformatics*, vol. 6, p. 165, 2005. 46

[241] B. Ryu, J. Jones, N. J. Blades, G. Parmigiani, M. A. Hollingsworth, R. H. Hruban, and S. E. Kern, "Relationships and differentially expressed genes among pancreatic cancers examined by large-scale serial analysis of gene expression.," *Cancer research*, vol. 62, pp. 819–826, Feb. 2002. 46

[242] S. Anders and W. Huber, "Differential expression analysis for sequence count data.," *Genome biology*, vol. 11, no. 10, p. R106, 2010. 47

[243] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 136–138, Jan. 2010. 47

[244] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.," *Bioinformatics (Oxford, England)*, vol. 19, pp. 185–193, Jan. 2003.

[245] Y. Yang, S. Dudoit, P. Luu, and D. Lin, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic acids ...*, 2002. 47

[246] T. J. Hardcastle and K. A. Kelly, "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.," *BMC bioinformatics*, vol. 11, p. 422, 2010. 47

[247] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments.," *Statistical applications in genetics and molecular biology*, vol. 3, p. Article3, 2004. 47

[248] 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*, vol. 60, pp. 774–782, Sept. 2004. 49

[249] Y. Benjamini, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B ...*, 1995. 49, 50

[250] Y. Benjamini, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, 2001. 49, 50

[251] A. Reiner, D. Yekutieli, and Y. Benjamini, "Identifying differentially expressed genes using false discovery rate controlling procedures.," *Bioinformatics (Oxford, England)*, vol. 19, pp. 368–375, Feb. 2003. 50

[252] Y. Benjamini, "Quantitative trait loci analysis using the false discovery rate," *Genetics*, 2005. 50

[253] C. Llave, Z. Xie, K. D. Kasschau, and J. C. Carrington, "Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA.," *Science (New York, N.Y.)*, vol. 297, pp. 2053–2056, Sept. 2002. 50

[254] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel, "Prediction of plant microRNA targets.," *Cell*, vol. 110, pp. 513–520, Aug. 2002.

[255] G. Tang, B. J. Reinhart, D. P. Bartel, and P. D. Zamore, "A biochemical framework for RNA silencing in plants.," *Genes & development*, vol. 17, pp. 49–63, Jan. 2003. 50

[256] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions.," *Cell*, vol. 136, pp. 215–233, Jan. 2009. 50, 51, 52, 53, 121

[257] B. Lewis and C. Burge, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, 2005. 50, 52, 54, 106

[258] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel, "The impact of microRNAs on protein output.," *Nature*, vol. 455, pp. 64–71, Sept. 2008. 50

[259] C. B. Nielsen, N. Shomron, R. Sandberg, E. Hornstein, J. Kitzman, and C. B. Burge, "Determinants of targeting by endogenous and exogenous microRNAs and siRNAs.," *Rna*, vol. 13, pp. 1894–1910, Nov. 2007. 50

[260] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel, "MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing," *Molecular cell*, vol. 27, pp. 91–105, July 2007. 50, 52, 54

[261] A. C. Mallory, B. J. Reinhart, M. W. Jones-Rhoades, G. Tang, P. D. Zamore, M. K. Barton, and D. P. Bartel, "MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region.," *The EMBO journal*, vol. 23, pp. 3356–3364, Aug. 2004. 51

[262] K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel, "The widespread impact of mammalian MicroRNAs on mRNA repression and evolution.," *Science (New York, N.Y.)*, vol. 310, pp. 1817–1821, Dec. 2005. 52

[263] K. C. Miranda, T. Huynh, Y. Tay, Y.-S. Ang, W.-L. Tam, A. M. Thomson, B. Lim, and I. Rigoutsos, "A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes," *Cell*, vol. 126, pp. 1203–1217, Sept. 2006. 52

[264] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microRNA target recognition," *Nature genetics*, vol. 39, pp. 1278–1284, Sept. 2007. 52

[265] B. Lewis, I. Shih, M. Jones-Rhoades, and D. Bartel, "Prediction of mammalian microRNA targets," *Cell*, 2003. 52, 54

[266] I. L. Hofacker, "How microRNAs choose their targets," *Nature genetics*, vol. 39, pp. 1191–1192, Oct. 2007. 52

[267] D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan, "Inference of miRNA targets using evolutionary conservation and pathway analysis.," *BMC bioinformatics*, vol. 8, p. 69, 2007. 52

[268] P. Saetrom, B. S. E. Heale, O. Snøve, L. Aagaard, J. Alluin, and J. J. Rossi, "Distance constraints between microRNA target sites dictate efficacy and cooperativity.," *Nucleic acids research*, vol. 35, no. 7, pp. 2333–2342, 2007. 52

[269] D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander, "The microRNA.org resource: targets and expression," *Nucleic acids research*, vol. 36, pp. D149–D153, Dec. 2007. 53, 54, 66, 106

[270] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human MicroRNA targets.," *PLoS biology*, vol. 2, p. e363, Nov. 2004. 53, 66

[271] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster, "Complete suboptimal folding of RNA and the stability of secondary structures.," *Biopolymers*, vol. 49, pp. 145–165, Feb. 1999. 53

[272] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.," *Genome research*, vol. 15, pp. 1034–1050, Aug. 2005. 54, 56

[273] D. Betel, A. Koppal, P. Agius, and C. Sander, "Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites," *Genome biology*, 2010. 54

[274] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky, "Combinatorial microRNA target predictions.," *Nature genetics*, vol. 37, pp. 495–500, May 2005. 54, 106

[275] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich, "Fast and effective prediction of microRNA/target duplexes.," *Rna*, vol. 10, pp. 1507–1517, Oct. 2004. 54

[276] S. Lall, D. Grün, A. Krek, K. Chen, Y.-L. Wang, C. N. Dewey, P. Sood, T. Colombo, N. Bray, P. MacMenamin, H.-L. Kao, K. C. Gunsalus, L. Pachter, F. Piano, and N. Rajewsky, "A genome-wide map of conserved microRNA targets in C. elegans.," *Current Biology*, vol. 16, pp. 460–471, Mar. 2006. 54

[277] D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding, "VarScan: variant detection in massively parallel sequencing of individual and pooled samples.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2283–2285, Sept. 2009. 55, 56, 72

[278] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T.

Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors.," *Nature*, vol. 437, pp. 376–380, Sept. 2005. 55, 57

[279] D. A. Benson, I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank.," *Nucleic acids research*, Dec. 2011. 56

[280] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation.," *Nucleic acids research*, vol. 29, pp. 308–311, Jan. 2001. 56, 74

[281] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "McKusick's Online Mendelian Inheritance in Man (OMIM).," *Nucleic acids research*, vol. 37, pp. D793–6, Jan. 2009. 56, 74

[282] R. Apweiler, A. Bairoch, and C. Wu, "UniProt: the universal protein knowledgebase," *Nucleic acids ...*, 2004. 56

[283] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations.," *Nature methods*, vol. 7, pp. 248–249, Apr. 2010. 57, 72

[284] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.," *Nature Protocols*, vol. 4, no. 7, pp. 1073–1081, 2009. 57, 72

[285] D. Koboldt, L. Ding, and E. Mardis, "Challenges of sequencing human genomes," *Briefings in ...*, 2010. 57, 119

[286] G. T. Marth, F. Yu, A. R. Indap, K. Garimella, S. Gravel, W. F. Leong, C. Tyler-Smith, M. Bainbridge, T. Blackwell, X. Zheng-Bradley, Y. Chen, D. Challis, L. Clarke, E. V. Ball, K. Cibulskis, D. N. Cooper, B. Fulton, C. Hartl, D. Koboldt, D. Muzny, R. Smith, C. Sougnez, C. Stewart, A. Ward, J. Yu, Y. Xue, D. Altshuler, C. D. Bustamante, A. G. Clark, M. Daly, M. Depristo, P. Flicek, S. Gabriel, E. Mardis, A. Palotie, R. Gibbs, and the 1000 Genomes Project, "The functional spectrum of low-frequency coding variation.," *Genome biology*, vol. 12, p. R84, Sept. 2011. 58, 74, 78, 80

[287] Y. Li, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, H. Jiang, A. Albrechtsen, G. Andersen, H. Cao, T. Korneliussen, N. Grarup, Y. Guo, I. Hellman, X. Jin, Q. Li, J. Liu, X. Liu, T. Sparsø, M. Tang, H. Wu, R. Wu, C. Yu, H. Zheng, A. Astrup, L. Bolund, J. Holmkvist, T. Jørgensen, K. Kristiansen, O. Schmitz, T. W. Schwartz, X. Zhang, R. Li, H. Yang, J. Wang, T. Hansen, O. Pedersen, R. Nielsen, and J. Wang, "Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants.," *Nature genetics*, vol. 42, pp. 969–972, Nov. 2010. 58, 74, 78

[288] M. Latronico and D. Catalucci, "MicroRNA and cardiac pathologies," *Physiological ...*, 2008. 65, 114

[289] H. Li, B. Handsaker, A. Wysoker, and T. Fennell, "The sequence alignment/map format and SAMtools," ..., 2009. 72

[290] M. J. CORYDON and J. VOCKLEY, "Role of common gene variations in the molecular pathogenesis of short-chain acyl-CoA dehydrogenase deficiency," *Pediatric ...*, 2001. 83, 118

[291] S. Park, J.-S. Yang, Y.-E. Shin, J. Park, S. K. Jang, and S. Kim, "Protein localization as a principal feature of the etiology and comorbidity of genetic diseases.," *Molecular systems biology*, vol. 7, p. 494, May 2011. 83

[292] M. A. o. G. Expression, "Official website for Mouse Atlas of Gene Expression project," Feb. 2012. 87

[293] D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann, "RNA sequencing reveals two major classes of gene expression levels in metazoan cells.," *Molecular systems biology*, vol. 7, p. 497, 2011. 87, 91, 120

[294] J. Schlesinger, M. Tönjes, M. Schueler, Q. Zhang, I. Dunkel, and S. R. Sperling, "Evaluation of the LightCycler 1536 Instrument for high-throughput quantitative real-time PCR.," *Methods (San Diego, Calif.)*, vol. 50, pp. S19–22, Apr. 2010. 94

[295] S. W. Kong, Y. W. Hu, J. W. K. Ho, S. Ikeda, S. Polster, R. John, J. L. Hall, E. Bisping, B. Pieske, C. G. dos Remedios, and W. T. Pu, "Heart failure-associated changes in RNA splicing of sarcomere genes.," *Circulation. Cardiovascular genetics*, vol. 3, pp. 138–146, Apr. 2010. 95, 120

[296] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott, "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.," *Nucleic acids research*, vol. 40, pp. D130–5, Jan. 2012. 95

[297] P. Jääskeläinen, R. Miettinen, P. Kärkkäinen, L. Toivonen, M. Laakso, and J. Kuusisto, "Genetics of hypertrophic cardiomyopathy in eastern Finland: few founder mutations with benign or intermediary phenotypes.," *Annals of medicine*, vol. 36, no. 1, pp. 23–32, 2004. 95

[298] T. Klaavuniemi and J. Ylänne, "Zasp/Cypher internal ZM-motif containing fragments are sufficient to co-localize with alpha-actinin–analysis of patient mutations.," *Experimental cell research*, vol. 312, pp. 1299–1311, May 2006. 97

[299] E. van Rooij, L. B. Sutherland, J. E. Thatcher, J. M. DiMaio, R. H. Naseem, W. S. Marshall, J. A. Hill, and E. N. Olson, "Dysregulation of microRNAs after myocardial infarction reveals a role of miR-29 in cardiac fibrosis.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 13027–13032, Sept. 2008. 103

[300] H. Lu, R. J. Buchan, and S. A. Cook, "MicroRNA-223 regulates Glut4 expression and cardiomyocyte glucose metabolism.," *Cardiovascular research*, vol. 86, pp. 410–420, June 2010. 103

[301] E. Elvira-Matelot, X.-o. Zhou, N. Farman, G. Beaurain, A. Henrion-Caude, J. Hadchouel, and X. Jeunemaitre, "Regulation of WNK1 expression by miR-192 and aldosterone.," *Journal of the American Society of Nephrology : JASN*, vol. 21, pp. 1724–1731, Oct. 2010. 103

[302] J. Xie, T. Wu, K. Xu, I. K. Huang, O. Cleaver, and C.-L. Huang, "Endothelial-specific expression of WNK1 kinase is essential for angiogenesis and heart development in mice.," *The American journal of pathology*, vol. 175, pp. 1315–1327, Sept. 2009. 103

[303] I. Hofacker, W. Fontana, and P. Stadler, "Fast folding and comparison of RNA secondary structures," *Monatshefte für Chemie ...*, 1994. 104

[304] E. Bonnet, J. Wuyts, P. Rouzé, and Y. Van de Peer, "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.," *Bioinformatics (Oxford, England)*, vol. 20, pp. 2911–2917, Nov. 2004. 104

[305] B. C. Meyers, M. J. Axtell, B. Bartel, D. P. Bartel, D. Baulcombe, J. L. Bowman, X. Cao, J. C. Carrington, X. Chen, P. J. Green, S. Griffiths-Jones, S. E. Jacobsen, A. C. Mallory, R. A. Martienssen, R. S. Poethig, Y. Qi, H. Vaucheret, O. Voinnet, Y. Watanabe, D. Weigel, and J.-K. Zhu, "Criteria for annotation of plant MicroRNAs.," *The Plant cell*, vol. 20, pp. 3186–3190, Dec. 2008. 105

[306] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foà, J. Schliwka, U. Fuchs, A. Novosel, R.-U. Müller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H.-I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl, "A mammalian microRNA expression atlas based on small RNA library sequencing.," *Cell*, vol. 129, pp. 1401–1414, June 2007. 105

[307] D. Langenberger, C. Bermudez-Santana, J. Hertel, S. Hoffmann, P. Khaitovich, and P. F. Stadler, "Evidence for human microRNA-offset RNAs in small RNA sequencing data.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2298–2301, Sept. 2009. 105

[308] V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl, "A uniform system for microRNA annotation.," *Rna*, vol. 9, pp. 277–279, Mar. 2003. 105

[309] A. Chinchilla, E. Lozano, H. Daimi, F. J. Esteban, C. Crist, A. E. Aranega, and D. Franco, "MicroRNA profiling during mouse ventricular maturation: a role for miR-27 modulating Mef2c expression.," *Cardiovascular research*, vol. 89, pp. 98–108, Jan. 2011. 107

[310] A. Kuehbacher, C. Urbich, A. M. Zeiher, and S. Dimmeler, "Role of Dicer and Drosha for endothelial microRNA expression and angiogenesis.," *Circulation research*, vol. 101, pp. 59–68, July 2007. 107

[311] Y. Fukuda, H. Kawasaki, and K. Taira, "Exploration of human miRNA target genes in neuronal differentiation.," *Nucleic acids symposium series (2004)*, no. 49, pp. 341–342, 2005. 107

[312] K. Urbanek, M. C. Cabral-da Silva, N. Ide-Iwata, S. Maestroni, F. Delucchi, H. Zheng, J. Ferreira-Martins, B. Ogórek, D. D'Amario, M. Bauer, G. Zerbini, M. Rota, T. Hosoda, R. Liao, P. Anversa, J. Kajstura, and A. Leri, "Inhibition of notch1-dependent cardiomyogenesis leads to a dilated myopathy in the neonatal heart.," *Circulation research*, vol. 107, pp. 429–441, Aug. 2010. 107

[313] Y. Tsuchiya, M. Nakajima, S. Takagi, T. Taniya, and T. Yokoi, "MicroRNA regulates the expression of human cytochrome P450 1B1.," *Cancer research*, vol. 66, pp. 9090–9098, Sept. 2006. 107

[314] 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.)*, vol. 276, pp. 1404–1407, May 1997. 107

[315] F. Pichiorri, S.-S. Suh, M. Ladetto, M. Kuehl, T. Palumbo, D. Drandi, C. Taccioli, N. Zanesi, H. Alder, J. P. Hagan, R. Munker, S. Volinia, M. Boccadoro, R. Garzon, A. Palumbo, R. I. Aqeilan, and C. M. Croce, "MicroRNAs regulate critical genes associated with multiple myeloma pathogenesis.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, pp. 12885–12890, Sept. 2008. 107

[316] L. Yan, Q. Tang, D. Shen, S. Peng, Q. Zheng, H. Guo, M. Jiang, and W. Deng, "SOCS-1 inhibits TNF-alpha-induced cardiomyocyte apoptosis via ERK1/2 pathway activation.," *Inflammation*, vol. 31, pp. 180–188, June 2008. 108

[317] V. Olive, M. J. Bennett, J. C. Walker, C. Ma, I. Jiang, C. Cordon-Cardo, Q.-J. Li, S. W. Lowe, G. J. Hannon, and L. He, "miR-19 is a key oncogenic component of mir-17-92.," *Genes & development*, vol. 23, pp. 2839–2849, Dec. 2009. 108

[318] G. Y. Oudit and J. M. Penninger, "Cardiac regulation by phosphoinositide 3-kinases and PTEN.," *Cardiovascular research*, vol. 82, pp. 250–260, May 2009. 108

[319] W. Ye, Q. Lv, C.-K. A. Wong, S. Hu, C. Fu, Z. Hua, G. Cai, G. Li, B. B. Yang, and Y. Zhang, "The effect of central loops in miRNA:MRE duplexes on the efficiency of miRNA-mediated gene regulation.," *PloS one*, vol. 3, no. 3, p. e1719, 2008. 108

[320] T. Zhao, W. Zhao, Y. Chen, R. A. Ahokas, and Y. Sun, "Vascular endothelial growth factor (VEGF)-A: role on cardiac angiogenesis following myocardial infarction.," *Microvascular research*, vol. 80, pp. 188–194, Sept. 2010. 108

[321] L. Castellano, G. Giamas, J. Jacob, R. C. Coombes, W. Lucchesi, P. Thiruchelvam, G. Barton, L. R. Jiao, R. Wait, J. Waxman, G. J. Hannon, and J. Stebbing, "The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 15732–15737, Sept. 2009. 108

[322] V. Leibetseder, S. Humpeler, A. Zuckermann, M. Svoboda, T. Thalhammer, W. Marktl, and C. Ekmekcioglu, "Time dependence of estrogen receptor expression in human hearts.," *Biomedicine & pharmacotherapy = Biomédecine & pharmacothérapie*, vol. 64, pp. 154–159, Mar. 2010. 108

[323] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "miRecords: an integrated resource for microRNA-target interactions.," *Nucleic acids research*, vol. 37, pp. D105–10, Jan. 2009. 109

[324] P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou, "TarBase: A comprehensive database of experimentally supported animal microRNA targets.," *Rna*, vol. 12, pp. 192–197, Feb. 2006. 109

[325] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang, "miRTarBase: a database curates experimentally validated microRNA-target interactions.," *Nucleic acids research*, vol. 39, pp. D163–9, Jan. 2011. 109

[326] ENCODE Project Consortium, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.," *Nature*, vol. 447, pp. 799–816, June 2007. 113

[327] P. J. Farnham, "Insights from genomic profiling of transcription factors.," *Nature reviews. Genetics*, vol. 10, pp. 605–616, Sept. 2009. 113

[328] M. Lange, B. Kaynak, U. B. Forster, M. Tönjes, J. J. Fischer, C. Grimm, J. Schlesinger, S. Just, I. Dunkel, T. Krueger, S. Mebus, H. Lehrach, R. Lurz, J. Gobom, W. Rottbauer, S. Abdelilah-Seyfried, and S. Sperling, "Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex.," *Genes & development*, vol. 22, pp. 2370–2384, Sept. 2008. 113

[329] C. M. Koch, R. M. Andrews, P. Flicek, S. C. Dillon, U. Karaöz, G. K. Clelland, S. Wilcox, D. M. Beare, J. C. Fowler, P. Couttet, K. D. James, G. C. Lefebvre, A. W. Bruce, O. M. Dovey, P. D. Ellis, P. Dhami, C. F. Langford, Z. Weng, E. Birney, N. P. Carter, D. Vetrie, and I. Dunham, "The landscape of histone modifications across 1human cell lines.," *Genome research*, vol. 17, pp. 691–707, June 2007.

[330] A. J. Ruthenburg, H. Li, D. J. Patel, and C. D. Allis, "Multivalent engagement of chromatin modifications by linked binding modules.," *Nature reviews. Molecular cell biology*, vol. 8, pp. 983–994, Dec. 2007. 113

[331] Y. Wang, Y. Liang, and Q. Lu, "MicroRNA epigenetic alterations: predicting biomarkers and therapeutic targets in human diseases.," *Clinical genetics*, vol. 74, pp. 307–315, Oct. 2008. 113

[332] K. R. Cordes and D. Srivastava, "MicroRNA regulation of cardiovascular development.," *Circulation research*, vol. 104, pp. 724–732, Mar. 2009.

[333] A. Bonauer, G. Carmona, M. Iwasaki, M. Mione, M. Koyanagi, A. Fischer, J. Burchfield, H. Fox, C. Doebele, K. Ohtani, E. Chavakis, M. Potente, M. Tjwa, C. Urbich, A. M. Zeiher, and S. Dimmeler, "MicroRNA-92a controls angiogenesis and functional recovery of ischemic tissues in mice.," *Science (New York, N.Y.)*, vol. 324, pp. 1710–1713, June 2009. 113

[334] 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)*, vol. 25, pp. 1715–1721, July 2009. 114

[335] D. E. Schones and K. Zhao, "Genome-wide approaches to studying chromatin modifications.," *Nature reviews. Genetics*, vol. 9, pp. 179–191, Mar. 2008. 114

[336] V. N. Kim, J. Han, and M. C. Siomi, "Biogenesis of small RNAs in animals.," *Nature reviews. Molecular cell biology*, vol. 10, pp. 126–139, Feb. 2009. 115

[337] S. T. MacDonald, S. D. Bamforth, C.-M. Chen, C. R. Farthing, A. Franklyn, C. Broadbent, J. E. Schneider, Y. Saga, M. Lewandoski, and S. Bhattacharya, "Epiblastic Cited2 deficiency results in cardiac phenotypic heterogeneity and provides a mechanism for haploinsufficiency.," *Cardiovascular research*, vol. 79, pp. 448–457, Aug. 2008. 115

[338] M. Schueler, Q. Zhang, J. Schlesinger, M. Tönjes, and S. R. Sperling, "Dynamics of Srf, p300 and histone modifications during cardiac maturation in mouse.," *Molecular bioSystems*, vol. 8, pp. 495–503, Feb. 2012. 115

[339] H. Siomi and M. C. Siomi, "On the road to reading the RNA-interference code.," *Nature*, vol. 457, pp. 396–404, Jan. 2009. 115

[340] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel, "Global and local architecture of the mammalian microRNA-transcription factor regulatory network.," *PLoS computational biology*, vol. 3, p. e131, July 2007. 115

[341] E. van Rooij and E. N. Olson, "Searching for miR-acles in cardiac fibrosis.," *Circulation research*, vol. 104, pp. 138–140, Jan. 2009. 116

[342] A. Carè, D. Catalucci, F. Felicetti, D. Bonci, A. Addario, P. Gallo, M.-L. Bang, P. Segnalini, Y. Gu, N. D. Dalton, L. Elia, M. V. G. Latronico, M. Høydal, C. Autore, M. A. Russo, G. W. Dorn, O. Ellingsen, P. Ruiz-Lozano, K. L. Peterson, C. M. Croce, C. Peschle, and G. Condorelli, "MicroRNA-133 controls cardiac hypertrophy.," *Nature medicine*, vol. 13, pp. 613–618, May 2007.

[343] T. Thum, C. Gross, J. Fiedler, T. Fischer, S. Kissler, M. Bussen, P. Galuppo, S. Just, W. Rottbauer, S. Frantz, M. Castoldi, J. Soutschek, V. Koteliansky, A. Rosenwald, M. A. Basson, J. D. Licht, J. T. R. Pena, S. H. Rouhanifard, M. U. Muckenthaler, T. Tuschl, G. R. Martin, J. Bauersachs, and S. Engelhardt, "MicroRNA-21 contributes to myocardial disease by stimulating MAP kinase signalling in fibroblasts.," *Nature*, vol. 456, pp. 980–984, Dec. 2008.

[344] E. van Rooij, L. B. Sutherland, N. Liu, A. H. Williams, J. McAnally, R. D. Gerard, J. A. Richardson, and E. N. Olson, "A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 18255–18260, Nov. 2006. 121

[345] E. van Rooij, L. B. Sutherland, X. Qi, J. A. Richardson, J. Hill, and E. N. Olson, "Control of stress-dependent cardiac growth and gene expression by a microRNA.," *Science (New York, N.Y.)*, vol. 316, pp. 575–579, Apr. 2007.

[346] B. Yang, H. Lin, J. Xiao, Y. Lu, X. Luo, B. Li, Y. Zhang, C. Xu, Y. Bai, H. Wang, G. Chen, and Z. Wang, "The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2.," *Nature medicine*, vol. 13, pp. 486–491, Apr. 2007.

[347] V. Divakaran and D. L. Mann, "The emerging role of microRNAs in cardiac remodeling and heart failure.," *Circulation research*, vol. 103, pp. 1072–1083, Nov. 2008. 116, 121

[348] F. Cordero, M. Beccuti, M. Arigoni, S. Donatelli, and R. A. Calogero, "Optimizing a Massive Parallel Sequencing Workflow for Quantitative miRNA Expression Analysis.," *PloS one*, vol. 7, no. 2, p. e31630, 2012. 116

[349] G. P. Sykiotis, L. Plummer, V. A. Hughes, M. Au, S. Durrani, S. Nayak-Young, A. A. Dwyer, R. Quinton, J. E. Hall, J. F. Gusella, S. B. Seminara, W. F. Crowley, and N. Pitteloud, "Oligogenic basis of isolated gonadotropin-releasing hormone deficiency.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 15140–15144, Aug. 2010. 117

[350] N. Katsanis, "The oligogenic properties of Bardet-Biedl syndrome.," *Human molecular genetics*, vol. 13 Spec No 1, pp. R65–71, Apr. 2004.

[351] N. D. E. Greene, P. Stanier, and A. J. Copp, "Genetics of human neural tube defects.," *Human molecular genetics*, vol. 18, pp. R113–29, Oct. 2009. 117

[352] J. Coakley, "PHYSICIAN's GUIDE TO THE LABORATORY DIAGNOSIS OF METABOLIC DISEASES," *Journal of Paediatrics and Child Health*, vol. 39, pp. 641–641, Nov. 2003. 118

[353] A. Freiburg and M. Gautel, "A molecular map of the interactions between titin and myosin-binding protein C. Implications for sarcomeric assembly in familial hypertrophic cardiomyopathy.," *European journal of biochemistry / FEBS*, vol. 235, pp. 317–323, Jan. 1996.

[354] N. Gregersen, B. S. Andresen, M. J. Corydon, T. J. Corydon, R. K. Olsen, L. Bolund, and P. Bross, "Mutation analysis in mitochondrial fatty acid oxidation defects: Exemplified by acyl-CoA dehydrogenase deficiencies, with special focus on genotype-phenotype relationship.," *Human mutation*, vol. 18, pp. 169–189, Sept. 2001. 118

[355] C. Pedersen, S. Kølvraa, A. Kølvraa, and V. Stenbroen, "... ACADS gene variation spectrum in 114 patients with short-chain acyl-CoA dehydrogenase (SCAD) deficiency is dominated by missense variations leading to protein ...," *Human genetics*, 2008. 118

[356] M. Satoh, M. Takahashi, T. Sakamoto, M. Hiroe, F. Marumo, and A. Kimura, "Structural analysis of the titin gene in hypertrophic cardiomyopathy: identification of a novel disease gene.," *Biochemical and biophysical research communications*, vol. 262, pp. 411–417, Aug. 1999. 118

[357] B. Gerull, M. Gramlich, J. Atherton, M. McNabb, K. Trombitás, S. Sasse-Klaassen, J. G. Seidman, C. Seidman, H. Granzier, S. Labeit, M. Frenneaux, and L. Thierfelder, "Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy.," *Nature genetics*, vol. 30, pp. 201–204, Feb. 2002. 118

[358] G. E. Davies, C. M. Howard, M. J. Farrer, M. M. Coleman, L. B. Bennett, L. M. Cullen, R. K. Wyse, J. Burn, R. Williamson, and A. M. Kessling, "Genetic variation in the COL6A1 region is associated with congenital heart defects in trisomy 21 (Down's syndrome).," *Annals of human genetics*, vol. 59, pp. 253–269, July 1995. 118

[359] T. R. Grossman, A. Gamliel, R. J. Wessells, O. Taghli-Lamallem, K. Jepsen, K. Ocorr, J. R. Korenberg, K. L. Peterson, M. G. Rosenfeld, R. Bodmer, and E. Bier, "Overexpression of DSCAM and COL6A2 cooperatively generates congenital heart defects.," *PLoS genetics*, vol. 7, p. e1002344, Nov. 2011. 118

[360] G. Michielon, B. Marino, R. Formigari, G. Gargiulo, F. Picchio, M. C. Digilio, S. Anaclerio, G. Oricchio, S. P. Sanders, and R. M. Di Donato, "Genetic syndromes and outcome

after surgical correction of tetralogy of Fallot.," *The Annals of thoracic surgery*, vol. 81, pp. 968–975, Mar. 2006. 118

[361] M. Nadeau, R. O. Georges, B. Laforest, A. Yamak, C. Lefebvre, J. Beauregard, P. Paradis, B. G. Bruneau, G. Andelfinger, and M. Nemer, "An endocardial pathway involving Tbx5, Gata4, and Nos3 required for atrial septum formation.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 19356–19361, Nov. 2010. 119

[362] R. B. H. Williams, E. K. F. Chan, M. J. Cowley, and P. F. R. Little, "The influence of genetic variation on gene expression.," *Genome research*, vol. 17, pp. 1707–1716, Dec. 2007. 119

[363] D. Ramsköld, E. T. Wang, C. B. Burge, and R. Sandberg, "An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.," *PLoS computational biology*, vol. 5, p. e1000598, Dec. 2009. 120

[364] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, "mRNA-Seq whole-transcriptome analysis of a single cell.," *Nature methods*, vol. 6, pp. 377–382, May 2009. 120

[365] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.," *Nat Biotechnol*, vol. 28, pp. 511–515, May 2010. 120

[366] A. G. L. Douglas and M. J. A. Wood, "RNA splicing: disease and therapy.," *Briefings in functional genomics*, vol. 10, pp. 151–164, May 2011. 120

[367] M. Nissim-Rafinia and B. Kerem, "Splicing regulation as a potential genetic modifier.," *Trends in genetics : TIG*, vol. 18, pp. 123–127, Mar. 2002. 120

[368] S. Waldmüller, S. Sakthivel, A. V. Saadi, C. Selignow, P. G. Rakesh, M. Golubenko, P. K. Joseph, R. Padmakumar, P. Richard, K. Schwartz, J. M. Tharakan, C. Rajamanickam, and H. P. Vosberg, "Novel deletions in MYH7 and MYBPC3 identified in Indian families with familial hypertrophic cardiomyopathy.," *Journal of molecular and cellular cardiology*, vol. 35, pp. 623–636, June 2003. 120

[369] N. Ohsawa, M. Koebis, S. Suo, I. Nishino, and S. Ishiura, "Alternative splicing of PDLIM3/ALP, for alpha-actinin-associated LIM protein 3, is aberrant in persons with myotonic dystrophy.," *Biochemical and biophysical research communications*, vol. 409, pp. 64–69, May 2011. 120

[370] P. Pomiès, M. Pashmforoush, C. Vegezzi, K. R. Chien, C. Auffray, and M. C. Beckerle, "The cytoskeleton-associated PDZ-LIM protein, ALP, acts on serum response factor activity to regulate muscle differentiation.," *Molecular biology of the cell*, vol. 18, pp. 1723–1733, May 2007. 120

[371] T. Thum, P. Galuppo, C. Wolf, J. Fiedler, S. Kneitz, L. W. van Laake, P. A. Doevendans, C. L. Mummery, J. Borlak, A. Haverich, C. Gross, S. Engelhardt, G. Ertl, and J. Bauersachs, "MicroRNAs in the human heart: a clue to fetal gene reprogramming in heart failure.," *Circulation*, vol. 116, pp. 258–267, July 2007. 121

[372] S. Ikeda, S. W. Kong, J. Lu, E. Bisping, H. Zhang, P. D. Allen, T. R. Golub, B. Pieske, and W. T. Pu, "Altered microRNA expression in human heart disease.," *Physiological genomics*, vol. 31, pp. 367–373, Nov. 2007. 121

[373] D. B. Constam and E. J. Robertson, "SPC4/PACE4 regulates a TGFbeta signaling network during axis formation.," *Genes & development*, vol. 14, pp. 1146–1155, May 2000. 121

[374] S. G. Tevosian, A. E. Deconinck, M. Tanaka, M. Schinke, S. H. Litovsky, S. Izumo, Y. Fujiwara, and S. H. Orkin, "FOG-2, a cofactor for GATA transcription factors, is essential for heart morphogenesis and development of coronary vessels from epicardium.," *Cell*, vol. 101, pp. 729–739, June 2000. 122

[375] X. Fu, N. Fu, S. Guo, Z. Yan, Y. Xu, H. Hu, C. Menzel, W. Chen, Y. Li, R. Zeng, and P. Khaitovich, "Estimating accuracy of RNA-Seq and microarrays with proteomics.," *BMC genomics*, vol. 10, p. 161, 2009. 122

[376] T. Maier, M. Güell, and L. Serrano, "Correlation of mRNA and protein in complex biological samples.," *FEBS letters*, vol. 583, pp. 3966–3973, Dec. 2009. 122

[377] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mRNA abundance in yeast.," *Molecular and cellular biology*, vol. 19, pp. 1720–1730, Mar. 1999. 122

[378] C. Vogel, R. d. S. Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva, "Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line.," *Molecular systems biology*, vol. 6, p. 400, Aug. 2010. 122

[379] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte, "Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation.," *Nat Biotechnol*, vol. 25, pp. 117–124, Jan. 2007. 122

[380] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel, "An evolutionarily conserved mechanism for controlling the efficiency of protein translation.," *Cell*, vol. 141, pp. 344–354, Apr. 2010. 122

[381] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley, "Continuous base identification for single-molecule nanopore DNA sequencing.," *Nature nanotechnology*, vol. 4, pp. 265–270, Apr. 2009. 122

[382] E. E. Schadt, S. Turner, and A. Kasarskis, "A window into third-generation sequencing.," *Human molecular genetics*, vol. 19, pp. R227–40, Oct. 2010. 122

Abbreviations

BWT	Burrows wheeler transform
CC	Correlation coefficient
CDS	Coding sequence
CHD	Congenital heart disease
ChIP	Chromatin immunoprecipitation
ChIP-chip	ChIP followed by microarray analysis
ChIP-seq	ChIP followed by next-generation sequencing
CNV	Copy number variation
DNA	Deoxyribonucleic acid
d-TGA	Dextro-transposition of the great arteries
EM	Expectation-maximization algorithm
EST	Expressed sequence tag
FDR	False discovery rate
FHF	First heart field
GA	Genome analyzer (Illumina)
GS	Genome sequencer (Roche/454)
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HMM	Hidden Markov model
H3ac	Histone 3 acetylation at lysine 9 and lysine 14
InDel	Insertion and deletion
RNA	Ribonucleic acid
LV	Left ventricle
MAF	Minor allele frequency
MDS	Multi-dimensional scaling
mRNA	Messenger RNA
mRNA-seq	Next-generation sequencing of (m)RNAs

miRNA	MicroRNA
miRNA-seq	Next-generation sequencing of miRNAs
NGS	Next-generation sequencing
NH	Normal heart
PCA	Principle component analysis
PCC	Pearson correlation coefficient
PCR	Polymerase chain reaction
piRNA	Piwi-interacting RNA
POEM	Proportion estimation method
Pol II	RNA Polymerase II
pre-mRNA	Precursor mRNA
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
PWM	Position weight matrix
qPCR	Quantitative PCR
qRT-PCR	Quantitative real-time PCR
RISC	RNA-induced silencing complex
RNA-seq	Next-generation sequencing of (m)RNAs
RPKM	Reads per kilobase per million mapped reads
rRNA	Ribosomal RNA
RV	Right ventricle
SHF	Secondary heart field
siRNA	Short interfering RNA
Small RNA-seq	Next-generation sequencing of small RNAs (like miRNAs)
SNV	Single nucleotide variation
SNP	Single nucleotide polymorphism
TF	Transcription factor
TFBS	Transcription factor binding site
TI	Tricuspid insufficiency
TMM	Trimmed mean of M-values
tRNA	Transfer RNA
TSS	Transcription start site
UTR	Untranslated region

Zusammenfassung

Im Bereich der Genanalyse hat es in den vergangenen Jahren eine wesentliche Abkehr von der Anwendung der halbautomatisierten Sanger-Sequenzierung hin zur sogenannten Next-Generation-Sequenzierung (NGS) gegeben. Der Hauptvorteil dieser NGS-Methoden liegt vor allem in der Fähigkeit Millionen von DNS-Fragmenten in sehr kurzer Zeit zu sequenzieren. Insgesamt gibt es eine breite Palette von NGS-Anwendungen, die sich schnell weiterentwickeln, was die computergestützte Analyse der damit verbundenen Datenmengen sehr anspruchsvoll macht. In der Genexpressionsanalyse werden die früher herkömmlichen Microarrays mehr und mehr durch sequenzbasierte Methoden ersetzt, die kodierenden und nicht-kodierenden Transkripte ohne deren vorherige Kenntnis identifizieren und quantifizieren können. Die Sequenzierung eines ganzen Genoms oder bestimmter Sequenzen (gezielte Resequenzierung) ermöglicht die Identifizierung von genomischen Variationen auf einer breiten Basis.

Diese Dissertation beschäftigt sich mit den Herausforderungen, die sich im Zusammenhang mit der Anwendung von NGS-Technologien ergeben. Das beinhaltet die gezielte DNA-Resequenzierung, die Sequenzierung von exprimierten mRNAs (RNA-seq) und microRNAs (miRNA-seq) sowie die Identifizierung von Protein-DNA-Wechselwirkungen, wie Bindungsstellen für Transkriptionsfaktoren oder Histonmodifikationen (ChIP-seq). Die innerhalb der Arbeitsgruppe generierten sowie öffentlich verfügbaren, experimentellen Datensätze wurden verwendet, um neuartige, computergestützte Ansätze und Methoden der Bioinformatik für die Analyse von NGS-Datensätzen zu entwickeln und schließlich biologische Fragen hinsichtlich der Herzfunktion und -krankheit zu beantworten.

Eine erste Studie konzentriert sich auf die kombinatorische Regulation von kardialen, DNA-bindenden Transkriptionsfaktoren (ChIP-seq von Srf) beeinflusst von Histonmodifikationen (Histon 3 Acetylierung) und regulatorischen miRNAs (miRNA-seq). Wie in *PLoS Genetics* im Jahr 2011 veröffentlicht, haben diese verschiedenen reg-

ulierenden Ebenen von mRNA-Profilen ein hohes Maß an Wechselwirkung und das Potenzial sich gegenseitig zu modulieren. Zum Beispiel wird die Wirkung von Srf maßgeblich durch das gleichzeitige Auftreten von Histon 3 Acetylierungsmarkierungen beeinflusst. Darüber hinaus können 45% aller differentiell exprimierten mRNAs im Srf Knockdown durch die unterschiedliche Expression von microRNAs erklärt werden. Ungefähr die Hälfte aller differentiell exprimierten mRNAs wird durch andere sekundäre Effekte beeinflusst. Um daher ein vollständiges Bild des regulatorischen Transkriptionsnetzwerkes und der zugrundeliegenden Funktion von Kardiomyozyten (Herzmuskelzellen) zu erhalten, müssen die verschiedenen Modulatoren in Zusammenhang miteinander betrachtet werden. Im Rahmen dieser Studie wurde das Programm MicroRazerS entwickelt (veröffentlicht in *Bioinformatics* 2010). MicroRazerS ist optimiert für das Mappen kleiner RNA-Sequenzen, wie zum Beispiel microRNAs oder andere kleine nicht-codierende RNAs, zu einem Referenz-Genom. Es zeichnet sich durch eine höhere Sensitivität und zumindest vergleichbare Geschwindigkeit im Vergleich zu anderen Mapping-Programmen aus. Die Ergebnisse zeigen, dass MicroRazerS das Auffinden und die Entdeckung von microRNAs in Hochdurchsatz-Sequenzierungdaten wesentlich erleichtern kann.

Ein zweites Projekt zielte darauf ab, die genetische Grundlage der Fallot'schen Tetralogie (TOF) zu identifizieren. TOF tritt in bis zu 10% aller angeborenen Herzerkrankungen auf, die die größte Gruppe der angeborenen Fehlbildungen des Menschen darstellen. Diese Studie zeigt erstmals, dass TOF eine oligogenetische Erkrankung ist (Grunert *et al.* Manuskript unter Begutachtung). Wir haben eine mehrstufige Studie durchgeführt, darunter die gezielte Resequenzierung von über 1.000 herz- und muskelrelevanten Genen und microRNAs in TOF Patienten, Eltern und Kontrollen sowie die Analyse des ganzen Transkriptoms und miRNomes in TOF Patienten und gesunden Personen unter der Verwendung von NGS-Technologies (87 Proben). Gene wurden nach dem Vorhandensein von schädlichen Variationen und ihrer Mutationsrate in den TOF-Patienten im Vergleich zu gesunden Kontrollen (200 Fälle) beurteilt. Eine Menge von 16 sogenannten TOF-Genen wurde identifiziert, von denen durchschnittlich vier Gene pro TOF-Patient mutiert sind und die die TOF-Patienten von den Kontrollen unterscheiden. Im Allgemeinen stellt die in dieser Studie entwickelte Analysestrategie und der verwendete Bioinformatikansatz eine neue Perspektive für die Analyse von oligo- oder multigenetische Erkrankungen dar.

Summary

Over the past years, there has been a fundamental shift away from the application of semi-automated Sanger sequencing for genome analysis to so-called next-generation sequencing (NGS). The main advantage offered by NGS is the ability to sequence millions of DNA fragments in a very short time scale. There is a wide range of NGS applications, rapidly developing, making the computational analysis of their associated datasets very challenging. For gene expression analysis microarrays are more and more being replaced by sequenced-based methods, which can identify and quantify coding and non-coding transcripts without prior knowledge. Genome sequencing either at a whole or for particular sequences (targeted resequencing) enable the identification of genomic variations at a broad scale.

This thesis approaches computational challenges of NGS technologies applied for targeted DNA resequencing, sequencing of expressed mRNAs (RNA-seq) and miRNAs (miRNA-seq) as well as the identification of protein-DNA interactions such as transcription factor binding sites or chromatin histone marks (ChIP-seq). Experimental datasets generated within the group as well as publicly available were used to develop novel computational approaches and bioinformatics tools for the analysis of NGS datasets and eventually answer biological questions regarding cardiac function and disease.

A first study is focused on the combinatorial regulation of cardiac DNA-binding transcription factors (ChIP-seq of Srf) influenced by histone modifications (histone 3 acetylation) and regulatory miRNAs (miRNA-seq). As published in *PLoS Genetics* in 2011 these different levels regulating mRNA profiles have a high degree of interdependency and the potential to modulate each other. For example the effect of Srf binding is significantly influenced by the co-occurrence of histone 3 acetylation marks. Furthermore, differential expression of miRNAs can explain 45% of all differentially expressed mRNAs in Srf knockdown and approximately 50% of differential expression is driven

by other secondary effects. Thus, to obtain a full picture of the regulatory transcription network underlying cardiomyocyte function, the different modulators need to be viewed in context to each other. Within this project the tool MicroRazerS was developed (published in *Bioinformatics* 2010). MicroRazerS is optimized for mapping small RNAs such as miRNAs or other small non-coding RNAs onto a reference genome. It is characterized by a higher sensitivity and an at least comparable speed to other short read mapping tools. The results suggest that MicroRazerS can substantially facilitate the profiling and discovery of miRNAs obtained from high-throughput sequencing.

A second project aimed to identify the genetic basis of Tetralogy of Fallot (TOF). TOF accounts for up to 10% of all congenital heart disease, which are the most common birth defect in human. This study shows first time that TOF is an oligogenic disorder (Grunert *et al.* manuscript under review). We performed a multilevel study including targeted resequencing of over 1,000 heart- and muscle-relevant genes and miRNAs in TOF cases, parents and controls as well as whole transcriptome and miRNome analysis in TOF cases and healthy unaffected individuals using NGS techniques (87 samples). Genes were assessed according to the presence of deleterious variations and their rate of mutation in TOF subjects compared to healthy controls (200 cases). A set of 16 TOF genes was identified of which on average four genes per TOF subject are mutated and which discriminate TOF cases from controls. The computational approach developed within this study opens a new perspective for the analysis of oligo- or multigenic disorders in general.

Appendix A - The Srf Transcription Network

miRNA	Genomic location	Strand	Srf ChIP-seq peak position
mmu-miR-1-1	chr2:180123753-180123829	+	chr2:180120064-180120390
mmu-miR-1190	chr12:102259883-102260003	-	chr12:102267550-102267699
mmu-miR-1-2	chr18:10785479-10785550	-	chr18:10787723-10787902
mmu-miR-125b-1	chr9:41390009-41390085	+	chr9:41390294-41390404
mmu-miR-1306	chr16:18284301-18284371	-	chr16:18289279-18289438
mmu-miR-133a-1	chr18:10782907-10782974	-	chr18:10787723-10787902
mmu-miR-143	chr18:61808850-61808912	-	chr18:61811989-61812308
mmu-miR-145	chr18:61807479-61807548	-	chr18:61811989-61812308
mmu-miR-150	chr7:52377127-52377191	+	chr7:52384390-52384574
mmu-miR-1903	chr8:130883141-130883220	+	chr8:130882606-130882799
mmu-miR-1905	chr3:88340223-88340304	-	chr3:88330193-88330429
mmu-miR-191	chr9:108470650-108470723	+	chr9:108469232-108469334
mmu-miR-1934	chr11:69476545-69476627	+	chr11:69475763-69475992; chr11:69476189-69476343
mmu-miR-1966	chr8:108139366-108139473	+	chr8:108146258-108146513
mmu-miR-1967	chr8:126546541-126546622	+	chr8:126545158-126545460
mmu-miR-208b	chr14:55594537-55594613	-	chr14:55585452-55585748; chr14:55587192-55587412
mmu-miR-210	chr7:148407283-148407392	-	chr7:148414495-148414619
mmu-miR-2133-1	chr6:3151217-3151307	+	chr6: 3151462-3151625
mmu-miR-219-1	chr17:34161928-34162037	-	chr17:34168530-34168632
mmu-miR-688	chr15:102502223-102502297	-	chr15:102501312-102501743
mmu-miR-715	chr17:39981081-39981190	+	chr17:39979928-39983732; chr17:39984647-39985880
mmu-miR-9-3	chr7:86650150-86650239	+	chr7:86641075-86641194

Table S1: MicroRNAs with Srf binding events. ChIP-seq analysis revealed 22 miRNAs with at least one Srf binding event within a genomic region of ±10kb. Srf-ChIP peaks and miRNA positions based on mouse genome NCBI v37 (mm9).

miRNA (mmu)	Reads in siNon	Reads in Srf-si1	Reads in Srf-si2	Norm siNon	Norm Srf-si1	Norm Srf-si2	Up[1]/ down[-1] (Srf-si1/ siNon)	Up[1]/ down[-1] (Srf-si2/ siNon)	P-value Srf-si1/ siNon	P-value Srf-si2/ siNon
let-7d	30071	36690	54076	54,1	56,9	94,9	1	1	1.86E-09	0
let-7f-1	54979	59720	52368	98,9	92,5	91,9	-1	-1	3.74E-28	1.02E-227
let-7f-2	424277	495423	463067	763,5	767,7	812,6	1	1	0,04	3.35E-173
miR-101a	13748	14011	11760	24,7	21,7	20,6	-1	-1	2.54E-26	7.81E-130
miR-107	22797	29471	42243	41	45,7	74,1	1	1	9.57E-33	0
miR-125a	5335	5487	5155	9,6	8,5	9,1	-1	-1	2.18E-09	4.30E-20
miR-125b-2	1866	1896	1864	3,4	2,9	3,3	-1	-1	0	1.96E-05
miR-140	56373	60849	51693	101,4	94,3	90,7	-1	-1	2.79E-34	0
miR-146b	950	767	938	1,7	1,2	1,7	-1	-1	6.37E-13	0
miR-148a	1108	1152	1109	2	1,8	2	-1	-1	0,03	0
miR-148b	770	750	773	1,4	1,2	1,4	-1	-1	0	0,01
miR-151	1820	1545	1686	3,3	2,4	3	-1	-1	1.33E-18	1.41E-10
miR-152	15320	15940	13205	27,6	24,7	23,2	-1	-1	4.01E-21	5.30E-138
miR-16-1	1670	1683	1653	3	2,6	2,9	-1	-1	0	1.87E-05
miR-16-2	1337	1410	1291	2,4	2,2	2,3	-1	-1	0,04	1.06E-05
miR-182	1133	641	1090	2	1	1,9	-1	-1	9.50E-49	3.94E-05
miR-186	2736	2262	2321	4,9	3,5	4,1	-1	-1	6.62E-32	6.56E-28
miR-192	2410	2408	2200	4,3	3,7	3,9	-1	-1	1.29E-06	2.40E-15
miR-1937b	202	181	172	0,4	0,3	0,3	-1	-1	0,04	0,01
miR-195	709	706	614	1,3	1,1	1,1	-1	-1	0,02	4.84E-07
miR-196b	445	281	419	0,8	0,4	0,7	-1	-1	5.66E-15	0,01
miR-208b	1278	1110	1103	2,3	1,7	1,9	-1	-1	1.43E-11	3.14E-12
miR-21	45059	44580	39651	81,1	69,1	69,6	-1	-1	1.56E-124	0
miR-2134-1	117	600	1456	0,2	0,9	2,6	1	1	8.52E-63	1.34E-251
miR-2134-2	139	622	1436	0,3	1	2,5	1	1	1.44E-57	2.06E-230
miR-2134-3	8692	9705	7480	15,6	15	13,1	-1	-1	0,03	1.57E-79
miR-2134-4	147	634	1490	0,3	1	2,6	1	1	1.03E-56	4.97E-237
miR-2143-1	71	124	121	0,1	0,2	0,2	1	1	0,03	0,03
miR-2144	2252	3044	2405	4,1	4,7	4,2	1	1	3.05E-07	0,01
miR-22	8820	9705	8391	15,9	15	14,7	-1	-1	0	5.32E-38
miR-221	30791	37198	31657	55,4	57,6	55,6	1	1	2.10E-06	6.18E-51
miR-25	29288	23629	27014	52,7	36,6	47,4	-1	-1	0	2.68E-162
miR-26a-1	6802	6950	6676	12,2	10,8	11,7	-1	-1	6.37E-13	1.47E-21
miR-26a-2	6801	6954	6684	12,2	10,8	11,7	-1	-1	8.68E-13	3.18E-21
miR-27a	4853	6208	4993	8,7	9,6	8,8	1	1	2.80E-06	1.09E-08
miR-27b	14113	14742	13559	25,4	22,8	23,8	-1	-1	3.30E-18	8.16E-55
miR-28	739	674	749	1,3	1	1,3	-1	-1	3.72E-05	0,03
miR-29b-1	506	499	417	0,9	0,8	0,7	-1	-1	0,04	9.90E-07
miR-29b-2	907	929	889	1,6	1,4	1,6	-1	-1	0,03	0
miR-29c	8699	9685	7773	15,7	15	13,6	-1	-1	0,02	3.16E-62
miR-30a	31715	29834	32326	57,1	46,2	56,7	-1	-1	4.39E-148	9.39E-60
miR-30e	10280	9463	10226	18,5	14,7	18	-1	-1	2.34E-58	1.67E-27
miR-361	266	123	151	0,5	0,2	0,3	-1	-1	1.76E-17	5.66E-12
miR-378	103389	103695	88749	186,1	160,7	155,7	-1	-1	7.41E-238	0
miR-499	96808	104700	87731	174,2	162,2	154	-1	-1	4.75E-55	0
miR-532	2095	1968	1859	3,8	3,1	3,3	-1	-1	1.23E-10	2.23E-16
miR-689-2	429	611	939	0,8	1	1,7	1	1	0,01	3.70E-28
miR-92a-2	702	611	524	1,3	1	0,9	-1	-1	1.26E-06	8.74E-14
miR-93	1122	904	1121	2	1,4	2	-1	-1	2.53E-15	0

Table S2: Significantly deregulated miRNAs in Srf knockdown. 42 miRNAs (49 loci) were differentially expressed in Srf knockdown compared to control (siNon). miRNA loci based on mouse genome NCBI v37 (mm9). Matched reads to hairpin miRNA sequence based on miRBase annotations (release 14.0). P-values based on Fisher's exact test with p-value less than 0.05 after adjustment for multiple testing using Benjamini and Hochberg method for controlling the FDR.

Appendix B - Studying Tetralogy of Fallot

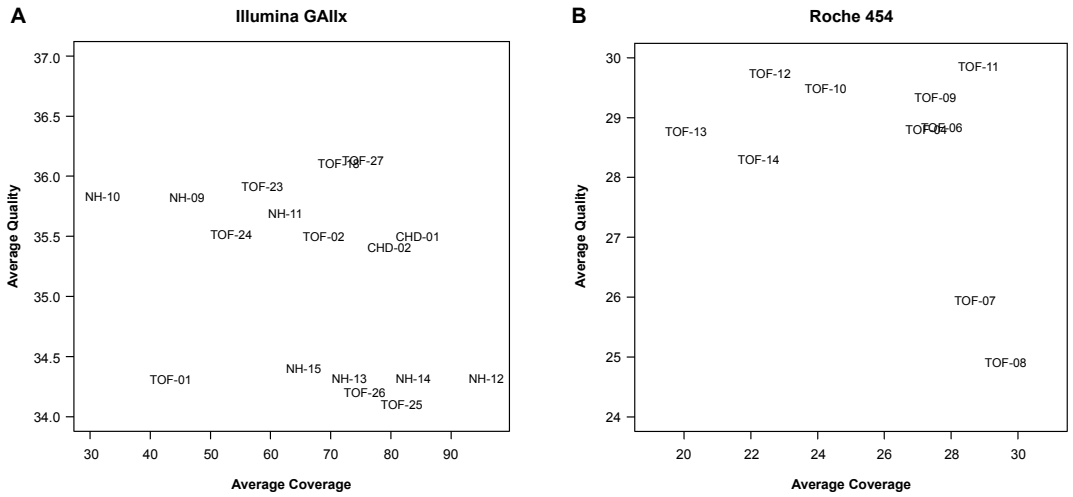

Figure S1: Scatterplot indicating average base quality (Phred scores) and coverage for samples measured using (A) Illumina Genome Analyzer IIx and (B) Roche/454 Genome Sequencer.

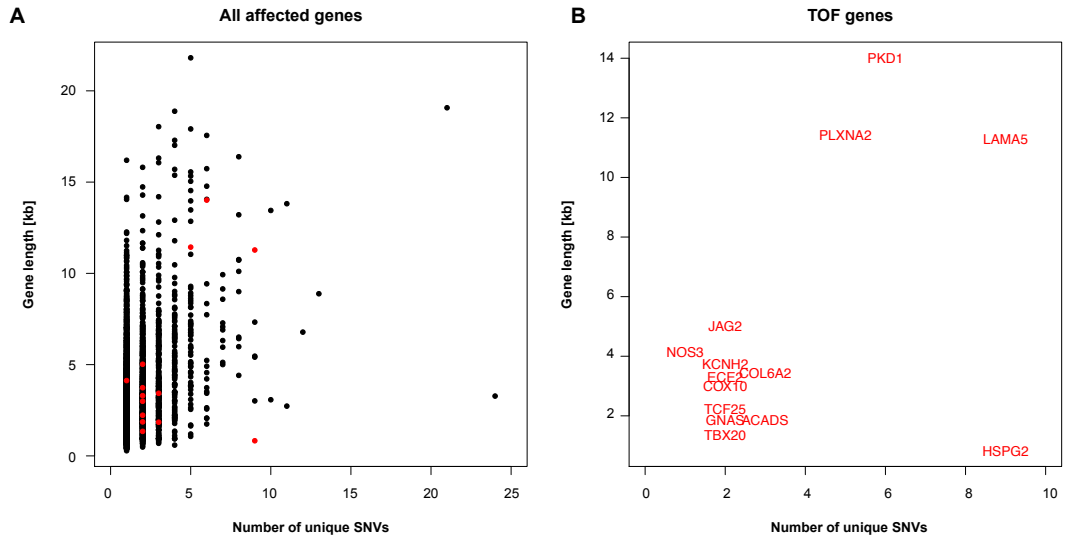

Figure S2: Quality control of found local variations. Scatterplot showing the number of found unique SNVs per gene against the gene (cDNA) lengths averaged over all transcripts for (A) all affected genes with the TOF genes marked in red and (B) only the 16 TOF genes. The number of SNVs per gene and the gene length shows no correlation, i.e. some short genes have a high number of unique SNVs while long genes can have only few SNVs. TTN (19 unique SNVs, ∼82 kb) was removed from the plot due to its length.

Gene	Genomic location	Patients	Notes
EXO1	chr1:240090785-240092165	TOF-07	1,379 bp deletion, located over one exon
FLII	chr17:18096625-18097629	TOF-07	1,003 bp deletion, located over 2 exons, cuts 10th exon and a small part of 9th exon
HCN2	chr19:556300-559434	TOF-04, TOF-06, TOF-07, TOF-12	3,134 bp deletion, cuts 4th exon and a large part of intron

Table S3: Identified copy number variations (CNVs) within the ten TOF samples pyrosequenced by the Roche/454 technology. Genomic locations based on NCBI v36.1 (hg18).

miRNA	Genomic location	Ref	Var	Gene	dbSNP ID	Sample ID(s)
miR-320b	chr1:222511382	-	AC	NVL	-	TOF-04
miR-412	chr14:100601607	A	G	-	rs61992671	TOF-01, TOF-02, TOF-07, TOF-12, TOF-13, TOF-14, TOF-18, TOF-25, TOF-26, TOF-27, NH-11, NH-12, NH-13, NH-15, CHD-01, CHD-02
miR-499-3p	chr20:33041912	A	G	MYH7B	rs3746444	TOF-06, TOF-09, TOF-14, NH-12, CHD-01
miR-532-3p	chrX:49654571	-	G	CLCN5	-	TOF-06, TOF-07, TOF-11

Table S4: Identified local variations in human mature miRNA sequences. Genomic locations based on NCBI v36.1 (hg18).

Sample	TOF-01	TOF-02	TOF-04	TOF-06	TOF-07	TOF-08	TOF-09	TOF-10	TOF-11	TOF-12	TOF-13	TOF-14	TOF-18	Average	%
SNVs with ≥1x RNA-seq reads	976	1245	1524	1366	1313	1249	1383	1491	1066	1326	1423	1009	430	1215	
- SNVs validated in RNA-seq (≥1x)	732	930	1126	1008	976	924	1041	1050	771	1027	1075	706	348	901	74%
SNVs with ≥5x RNA-seq reads	337	503	505	430	486	339	464	472	263	441	528	332	150	404	
- SNVs validated in RNA-seq (5x)	322	442	464	382	430	326	431	404	253	414	470	283	142	366	91%
SNVs with ≥10x RNA-seq reads	213	291	267	225	272	188	263	253	125	246	302	186	120	227	
- SNVs validated in RNA-seq (≥10x)	211	264	260	218	258	187	253	239	124	241	283	168	115	217	96%
INDELs with ≥1x RNA-seq reads	69	95	268	266	235	259	249	255	199	196	191	183	41	193	
- INDELs validated in RNA-seq (≥1x)	67	94	232	233	214	216	214	190	168	171	162	154	41	166	86%
INDELs with ≥5x RNA-seq reads	14	35	143	166	140	150	138	118	117	104	100	100	12	103	
- INDELs validated in RNA-seq (5x)	13	35	133	153	132	136	127	113	110	98	93	98	12	96	94%
INDELs with ≥10x RNA-seq reads	8	23	105	123	116	114	102	93	86	89	77	77	11	79	
- INDELs validated in RNA-seq (≥10x)	7	23	102	121	114	108	98	92	83	85	75	75	11	76	97%

Figure S3: Validation of local variations by RNA-seq reads.

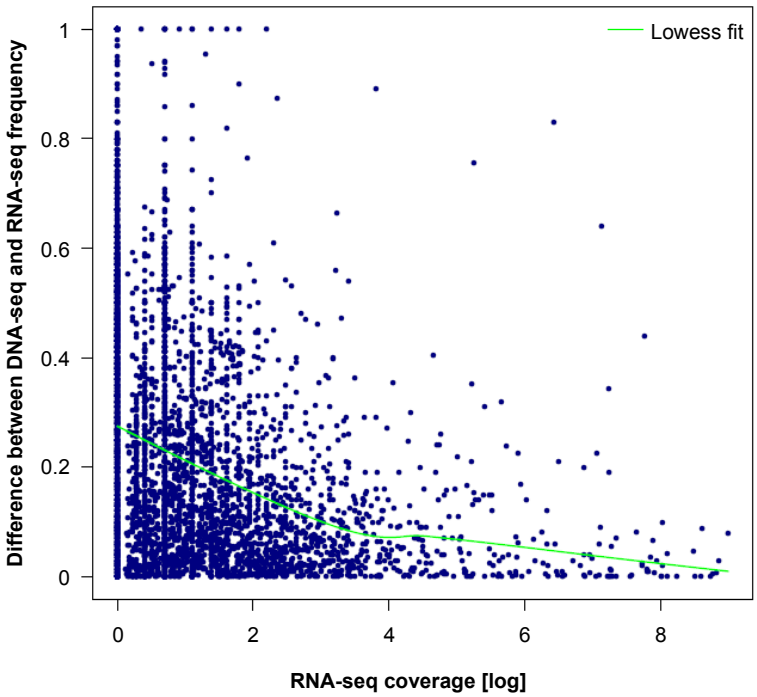

Figure S4: Verification of DNA sequencing (DNA-seq) results by RNA sequencing (RNA-seq). Scatterplot of the difference in local variance frequency measured by DNA-seq and RNA-seq dependent on the RNA-seq coverage. The higher the RNA-seq coverage the lower the distance between the two techniques. Data based on the average over all samples. The green line indicates a lowess fit of the data.

Gene	Genomic location	Ref	Var	dbSNP ID	Sample ID(s)
CACNA1B	chr9:139892491	A	T	-	TOF-27, NH-11, NH-12
CHFR	chr12:131959055	-	A	-	TOF-07, TOF-11
DYSF	chr2:71592561	T	G	-	TOF-02, TOF-18, TOF-24, NH-09, NH-12, NH-13, CHD-01
FLNA	chrX:153247868	A	C	-	TOF-01, TOF-23, NH-10, NH-12, NH-13, CHD-01
IA4	chr2:182082779	A	-	-	TOF-10
IL15	chr4:142870449	-	T	-	TOF-06
LAMB1	chr7:107363338	CT	AC	-	TOF-10, TOF-11, TOF-13, TOF-14
PAX8	chr2:113710770	T	C	-	TOF-26, NH-09, NH-12, NH-13, NH-15, CHD-01
PMM2	chr16:8849080	A	G	-	TOF-11
S100A13	chr1:151867220	-	C	-	TOF-26
SMYD1	chr2:88174191	T	C	-	TOF-07
SPEG	chr2:220050753	T	G	-	TOF-01, TOF-26
THRAP4	chr17:35433175	-	G	-	TOF-09
TTN	chr2:179122448	-	A	-	TOF-11

Table S5: Identified splice site mutations of high confidence local variations. Genomic locations based on NCBI v36.1 (hg18).

Appendix B

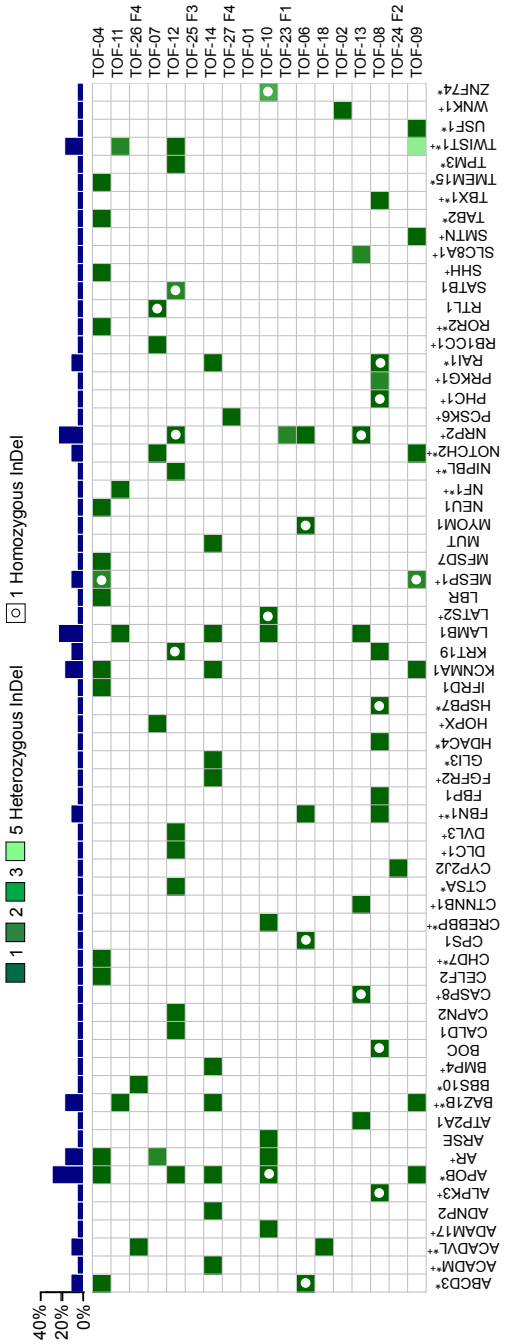

Figure S5: Distribution of InDels for affected genes with InDels only. Familial assignment is given after the sample identifier (F1 to F4). The number of InDels per gene is color-coded. Homozygous InDels are additionally marked by a white dot. Gene-wise frequencies of InDels are represented by blue bars. Gens marked with an asterisk have known associations with human disease affecting the heart, those marked with a cross show a cardiac phenotype when mutated or knocked out in mice.

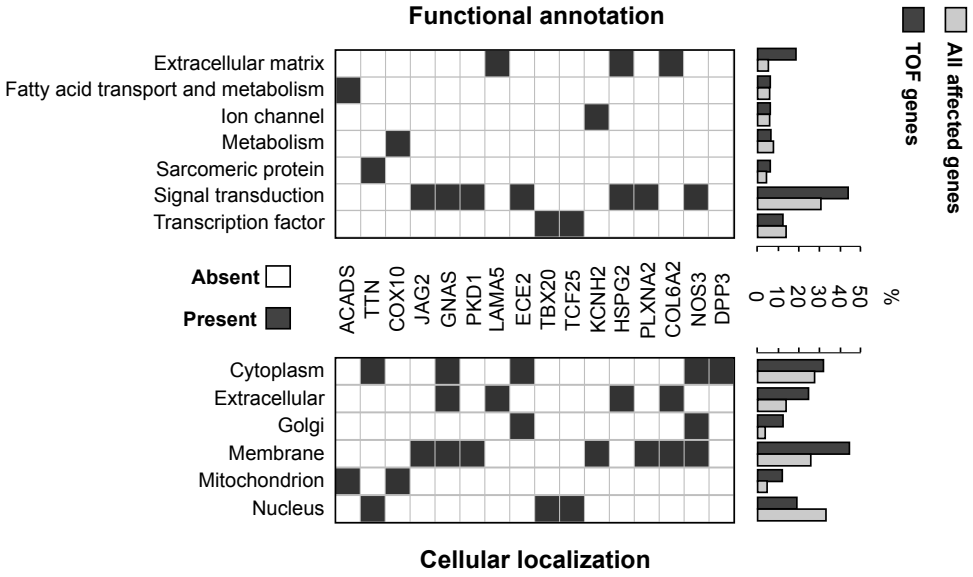

Figure S6: Heatmap of functional annotations (upper panel) and cellular localizations (lower panel) for the 16 affected TOF genes. Functional annotations have been assigned by literature curation. Cellular localizations were retrieved from the UniProt database. The frequency of each annotation in the 16 TOF genes (dark gray) as well as all affected genes (light gray) is shown on the right. "Signal transduction" and "Membrane" are clearly overrepresented in the TOF genes.

Appendix B

References for TOF genes.

	mouse development								human development				
Gene	E8.5	E9.5	E10.5	E11.5	E12.5	E13.5	E14.5	E15.5	week 7	week 8	week 9	week 10	PubMed ID
ACADS													
TTN	IHC	IHC											9440711; 2693040
COX10													
JAG2					ISH	IHC							17332426; 17273555
GNAS													
PKD1							BG						11593033
LAMA5					ISH	ISH							3256470
ECE2						ISH							10811845
TBX20	ISH	ISH	ISH	ISH	ISH								11118890
TCF25							ISH, NB						12107429
KCNH2			ISH		ISH								11557234
HSPG2		IHC		IHC			IHC	IHC					18694874; 10352025
PLXNA2					BG								19666519
COL6A2		IHC	IHC	IHC		IHC							9520112
NOS3		BG					BG						18556578
DPP3													

References for potential TOF genes.

	mouse development								human development				
Gene	E8.5	E9.5	E10.5	E11.5	E12.5	E13.5	E14.5	E15.5	week 7	week 8	week 9	week 10	PubMed ID
ALS2							BG						15686953
CACNA1C	PCR	PCR	PCR		PCR			PCR					12900400; 21079360
CACNA1H	PCR	ISH, PCR	PCR		ISH, PCR			ISH, PCR					12900400; 21079360; 12060068
CLTCL1													
MYOF								ISH					10607832
NOTCH1	ISH	ISH	IHC	PCR	ISH, PCR	PCR	PCR	PCR					17332426; 14701881; 12244553
SPEG				BG									19118250
ABCD3													
ACADVL									ISH		ISH		15845636
APOB													
AR										IHC	IHC	IHC	17968460
BAZ1B		ISH											19470456
FBN1	ISH												7829516
KCNMA1													
KRT19													
LAMB1										IHC	IHC		20552257
MESP1	ISH	BG											10393122; 11369261
NOTCH2					ISH								17273555
NRP2					ISH								11688557
RAI1													
TWIST1			ISH		ISH								22516205; 20804746

Figure S7: References for expression datasets. Published mRNA or protein expression data sets of TOF genes and potential TOF genes in developmental stages based on literature search. PCR: PCR or (quantitative) real-time PCR; ISH: *in situ* hybridisation; IHC: immunohistochemistry; BG: beta-galactosidase assay; NB: Northern Blot. CLTCL1 has no mouse homolog.

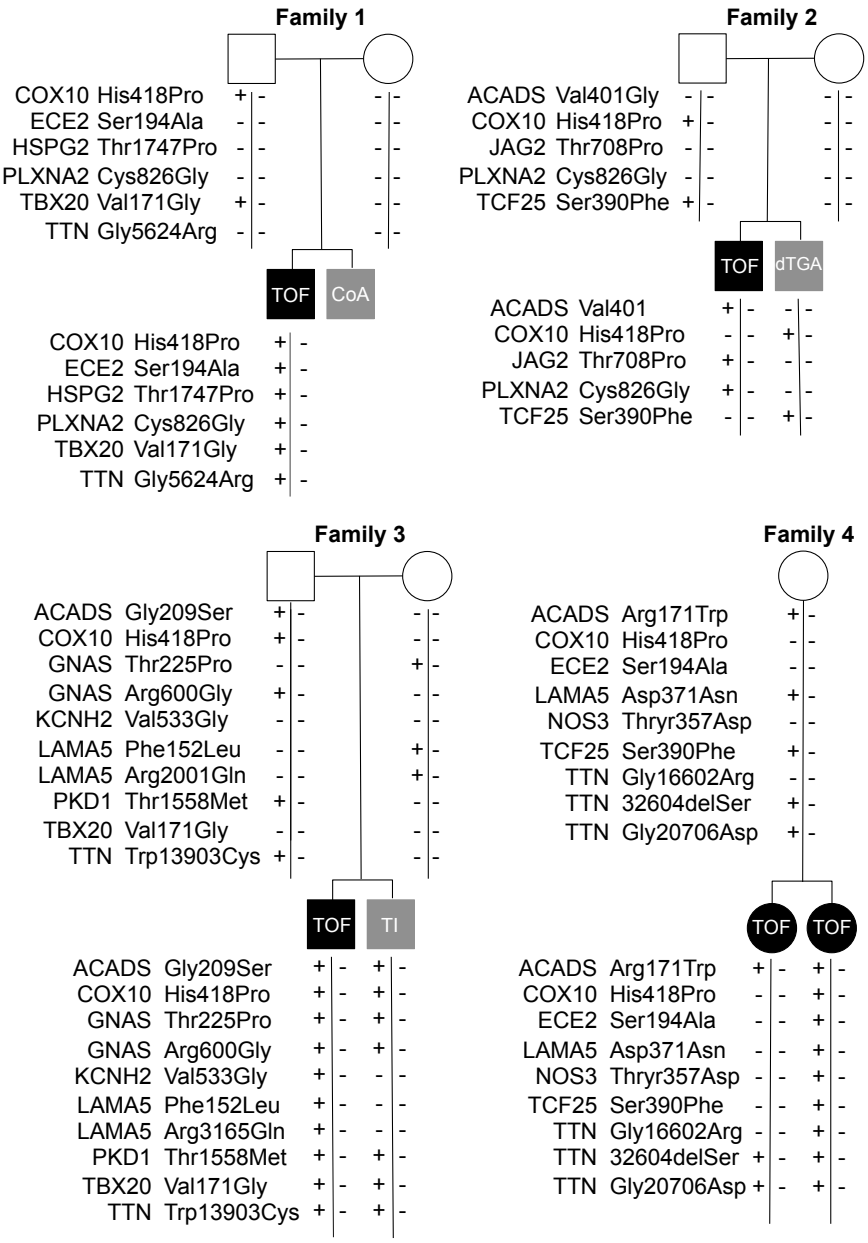

Figure S8: Pedigrees of the four analyzed families showing inherited and non-inherited local variations found in any of the 18 TOF patients.

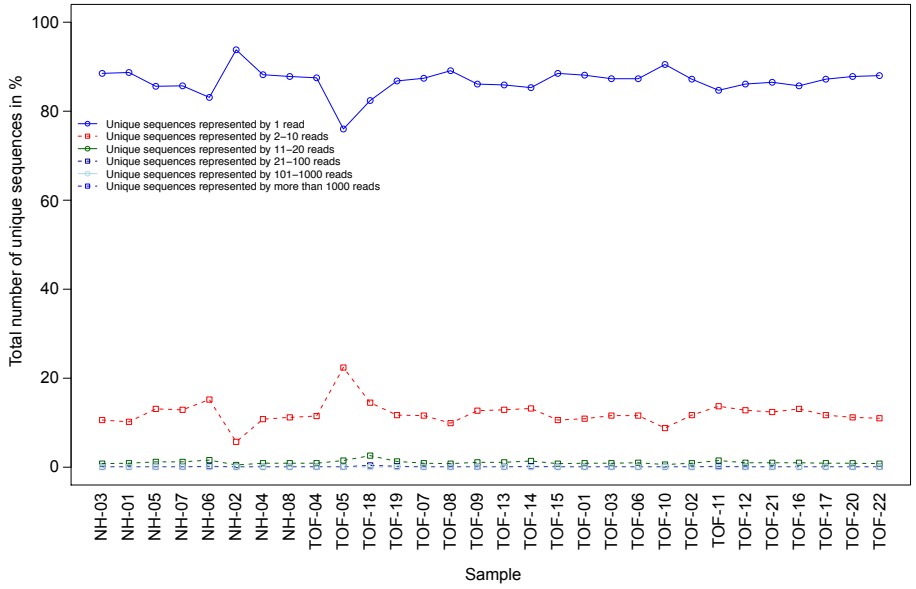

Figure S9: RNA-seq reads and their unique sequences obtained from mRNA libraries of patients with Tetralogy of Fallot (TOF) and healthy unaffected individuals (normal heart, NH).

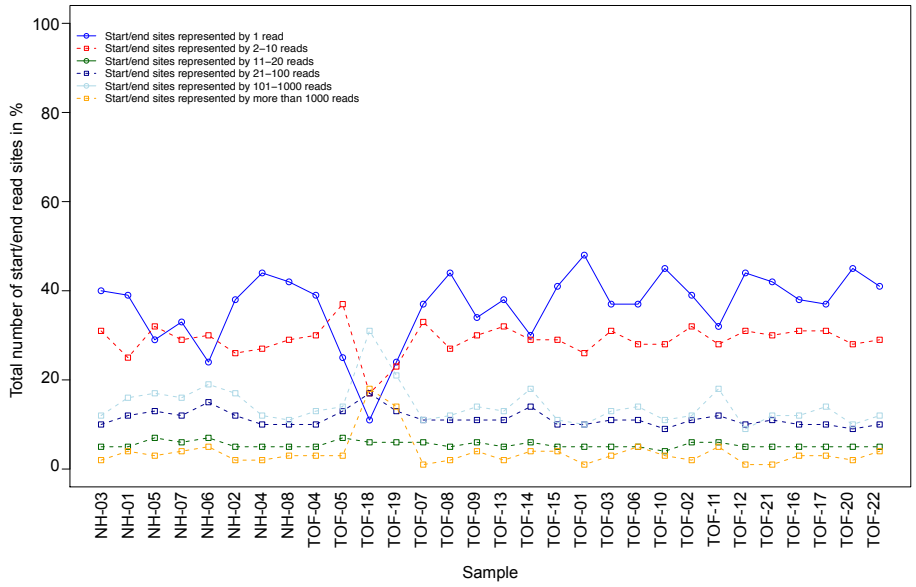

Figure S10: Perfectly identical pileups (reads with perfectly identical start/end sites) after unique mapping of RNA-seq reads obtained from mRNA libraries of patients with Tetralogy of Fallot (TOF) and healthy unaffected individuals (normal heart, NH).

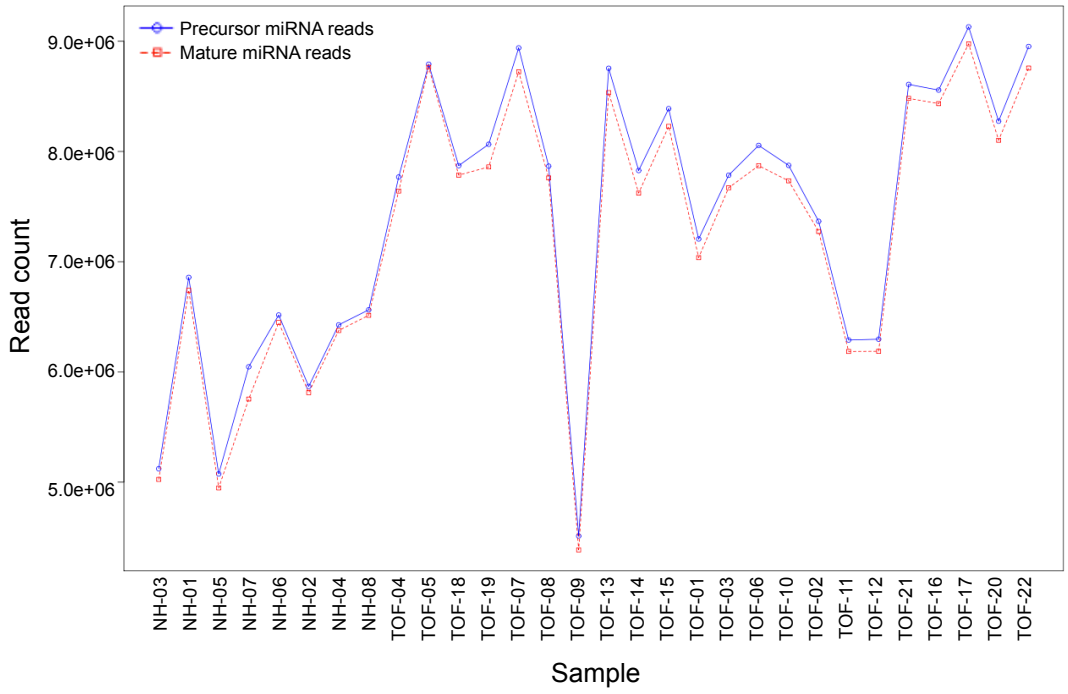

Figure S11: Mature and precursor miRNA read counts for TOF patients and healthy unaffected individuals (normal heart, NH).

Downregulated miRNAs
miR-10a, miR-29b, miR-29c, miR-98, miR-133b, miR-135a, miR-139-3p, miR-139-5p, miR-215, miR-1280,
Upregulated miRNAs
let-7b, let-7c, let-7i, miR-9, miR-15a, miR-15b, miR-17, miR-19b, miR-20a, miR-20b miR-23b, miR-26b, miR-27b, miR-28-3p, miR-30b, miR-32, miR-33a, miR-33b, miR-34a, miR-92a, miR-95, miR-101, miR-106a, miR-127-3p, miR-127-5p, miR-129-5p, miR-130a, miR-130b, miR-134, miR-136, miR-140-5p, miR-146a, miR-154, miR-181a, miR-181b, miR-181c, miR-181d, miR-186, miR-187, miR-192, miR-193a-5p, miR-204, miR-206, miR-210, miR-221, miR-222, miR-299-3p, miR-301a, miR-320a, miR-324-5p, miR-342-5p, miR-34c-5p, miR-361-5p, miR-362-5p, miR-363, miR-372, miR-376c, miR-378, miR-381, miR-382, miR-421, miR-422a, miR-423-3p, miR-424, miR-432, miR-433, miR-450a, miR-451, miR-452, miR-454, miR-455-5p, miR-499-3p, miR-504, miR-509-3-5p, miR-542-3p, miR-551b, miR-584 miR-590-5p, miR-618, miR-629, miR-651, miR-708, miR-769-5p, miR-886-5p, miR-887, miR-1185, miR-1246, miR-1259, miR-1261, miR-1262, miR-1285, miR-1287, miR-1977

Table S6: Significantly differentially expressed miRNAs ($p<0.05$) in TOF patients compared to right ventricle of healthy unaffected individuals.

Prediction ID	Mapped reads over all samples	Number of known miRNA precursors	Number of known human miRNA precursors	Well-formed secondary structure	Excision of mature seq (1=slightly imprecise, 2=precise)	3' Overhang in mature/ mature* duplex	Conservation (0=no, 1=poor, 2=good)	Offset RNAs	logFC TOF vs. RV	FDR TOF vs. RV	MirBase v15 annotation
chr6_705	207	18	1	yes	1	no	2	no	3.78	4.19E-006	
chr14_1457	61	0	0	yes	2	no	2	no	0.33	8.14E-001	miR-323b
chrX_1046	230	0	0	yes	2	yes	2	yes	0.29	7.67E-001	
chr1_2224	29	13	8	yes	2	no	1	no	0.00	1.00E+000	
chr3_203	18,268	11	2	yes	2	no	0	yes	5.33	9.32E-014	miR-378b
chr3_2417	30	3	0	yes	2	no	1	no	30.22	8.80E-006	
chr5_2052	28	0	0	no	2	no	2	yes	-2.68	2.27E-009	
chr6_2878	221	0	0	yes	1	no	1	no	1.09	1.10E-001	
chr9_1984	30	0	0	no	2	no	2	yes	1.95	3.08E-002	miR-219-2
chr11_1103	26	0	0	yes	2	no	1	no	0.00	1.00E+000	miR-3164
chr11_1180	29	0	0	yes	2	no	1	no	0.16	1.00E+000	
chr11_2198	406	3	0	yes	1	no	1	yes	0.00	1.00E+000	
chr16_1130	50,265	0	0	yes	2	no	0	yes	2.89	4.19E-006	
chr17_664	33	1	0	yes	2	no	1	yes	2.38	1.14E-002	
chr17_676	33	1	0	yes	2	no	1	yes	2.38	1.14E-002	
chr1_542	388	2	1	yes	2	no	0	yes	5.90	7.47E-012	
chr1_1097	479	0	0	yes	1	no	0	no	3.10	2.82E-006	miR-3117
chr1_3840	689	0	0	yes	1	no	0	no	0.67	3.05E-001	
chr2_1564	52	0	0	yes	2	no	0	no	2.85	2.34E-004	miR-3127
chr3_264	302	0	0	yes	2	no	0	no	1.92	3.36E-003	
chr3_3058	32	1	0	yes	2	no	0	no	0.00	1.00E+000	
chr5_1407	29	71	3	yes	2	no	0	no	-0.85	3.69E-002	
chr5_2051	28	0	0	yes	2	no	0	no	0.00	1.00E+000	
chr6_652	25	0	0	yes	2	no	0	no	6.69	6.87E-014	
chr10_2249	77	1	0	yes	?	no	0	yes	0.40	6.44E-001	miR-3158-2
chr11_393	26	34	0	yes	2	no	0	no	29.92	1.26E-005	
chr12_2214	25	1	0	yes	2	no	0	yes	0.00	1.00E+000	
chr22_318	94	0	0	yes	2	no	0	no	1.00	2.14E-001	
chrX_82	29	13	8	yes	2	no	0	no	0.00	1.00E+000	
chrX_885	61	0	0	yes	2	no	0	yes	1.30	9.38E-002	
chrX_1312	96	0	0	yes	2	no	0	no	2.55	2.70E-003	
chrX_1324	96	0	0	yes	2	no	0	no	2.43	2.10E-003	
chrX_2404	37	1	0	yes	2	no	0	yes	4.20	6.83E-005	

Figure S12: Novel miRNA candidates.

Curriculum Vitae

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

Curriculum Vitae

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

Curriculum Vitae

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

Curriculum Vitae

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

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.

Berlin, Juni 2012