Computational Analysis of High-Throughput Sequencing Data in Cardiac Disease and Skeletal Muscle Development

Vikas Bansal

February 2016

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

eingereicht im Fachbereich Mathematik und Informatik der Freien Universität Berlin

1. Betreuer: Prof. Dr. Martin Vingron

2. Betreuer: Prof. Dr. Silke Rickert-Sperling

Disputation: 21 July 2016

For my family, friends and research community.

Preface

The research described in the first part of the thesis (Chapter 4) was published in the journal *PLOS ONE* in 2014, under the title "Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with Tetralogy of Fallot" [1]. The research comprising the second part (Chapter 5), about epigenetic changes during myogenic differentiation, has not yet been published, but a manuscript describing an important regulatory mechanism to promote myogenic differentiation is in preparation. The last part of the thesis (Chapter 6) describes a pipeline to identify differential exon usage from RNA-seq data, and requires further investigation.

The full study in Chapter 4 describes a novel copy number variation calling method to identify individual disease-relevant copy number variations using exome or targeted resequencing data of small sets of samples. I contributed to this paper by developing the method, writing the code and testing the performance of the method. I evaluated the method using publicly available data of eight HapMap samples and subsequently applied it to a small number of Tetralogy of Fallot patients. Furthermore, I compared the method with tools published by others and was involved in writing the manuscript. For the study described in Chapter 5, I carried out the complete computational analysis of RNA-seq and ChIP-seq data. Moreover, I proposed a regulatory mechanism that might promote myogenic differentiation and furthermore, was involved in preparing the manuscript. For the last study (Chapter 6), I created a pipeline to identify differential exon usage from RNA-seq data meaning to identify the exons that are either excluded or included. Furthermore, I compared the results from our pipeline with a published method, Alternative Splicing Detector (ASD).

Acknowledgements

I want to express my deep and sincere gratitude to 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 supervisor Prof. Dr. Silke Rickert-Sperling for the opportunity to pursue this research and complete my PhD thesis in her group, for her patience, encouragement and opportunities to attend international conferences.

I would also like to thank Prof. Dr. Martin Vingron for supervising my PhD thesis and the opportunity to get involved in the meetings of Computational Molecular Biology Department at the MPIMG.

I thank Dr. Marcel Grunert for introducing me to the projects and helping me through the highs and lows of scientific research; Huanhuan Cui for cooperation and discussion on the Histone project; Kerstin Schulz, our "Chuck Norris", for all the discussions and coffee time; Ashley Cooper for editing this thesis; Dr. Markus Schüler, Dr. Cornelia Dorn, Dr. Sandra Schmitz, Dr. Elena Cano, Katherina Bellmann, Ilona Dunkel, Andreas Perrot, Sophia Schönhals and Sandra Appelt for your warm personalities and fruitful discussions; Martina Luig and Barbara Gibas for all the administrative assistance during my study.

I am deeply grateful to all patients and family members who generously participated in this research. I wish to thank all our collaborators and co-authors for their contributions. I would like to thank CardioNeT (Marie Curie Initial Training Network scheme).

I would like to express my further thanks to all my friends in Berlin, friends on Campus Buch, my Cricket team in Werder. Special thanks to my flatmate, Manvendra Singh and almost flatmate, Ankit Arora. Without your support, this work would have never been possible. Finally, my special thanks goes to my parents, my brothers, my sister-in-law for their love and understanding. And my nephew, Shivam Bansal, for giving me enthusiasm.

Contents

List of Figures

Chapter 1

Introduction

1.1 DNA and Gene Expression

Deoxyribonucleic acid (DNA) is a complex molecule that contains all of the genetic information necessary to build and maintain an organism. Nucleotide units attached to each other form a long stretch of DNA arranged in a double helix. Each nucleotide consists of three components: a nitrogenous base, a five-carbon sugar molecule (deoxyribose in the case of DNA) and a phosphate molecule. The backbone of the DNA is a chain of sugar and phosphate molecules. Each of the sugar groups are linked to one of the four nitrogenous bases i.e. cytosine (C), guanine (G), adenine (A) or thymine (T). In double-stranded DNA, A pairs with T and G pairs with C. In general, the genome size is defined as the total number of DNA base pairs in the haploid genome. For example, in humans there are about 3 billion base pairs per haploid genome. Eukaryotic cells package their genomic DNA into chromatin and arrange it in the cell nucleus as chromosomes. In humans, each somatic cell normally contains 23 pairs of chromosomes (22 pairs of autosomes and one pair of sex chromosomes). One set of 23 chromosomes is inherited from the father and the other from the mother.

Genes are the working subunits of DNA that are transcribed into RNA molecules, some of which (messenger RNA or mRNA) are translated into proteins [2]. Unlike double-stranded DNA, most RNA molecules are single-stranded and contain the unmethylated form of the base thymine called uracil (U). The synthesis of RNA from DNA, known as transcription, begins with the opening and unwinding of a small portion of the DNA double helix to expose the bases on each DNA strand. In eukaryotes, transcription of protein-coding genes is carried out by RNA polymerase II (Pol II). To initiate transcription, Pol II requires several initiation factors (general transcription factors), which escort and localize it to transcription start sites (TSS). For example,

general transcription factors like TFIIB or TFIID (complexes consisting of the TATA-binding protein and other associated factors) recruits Pol II to TSS, which is escorted by TFIIF [3]. In addition to the general transcription factors, sequence-specific DNA binding transcription factors (TFs) can govern gene transcription. These TFs bind to regulatory elements, which are generally found within several hundred or thousand bases of the start site of the gene. In general, regulatory elements like promoters and enhancers contain a fairly short DNA sequence (5 to 20 bp long), which is a specific binding site for one or more TFs [4]. TFs recognize these short sequences and bind to them to regulate the gene expression.

In eukaryotic cells, most genes are transcribed into precursor mRNA (pre-mRNA), which is processed to mature mRNA (or simply mRNA) and exported to the cytoplasm for translation [5]. This processing of pre-mRNA includes three major events: 5' capping, 3' polyadenylation, and RNA splicing. 5' capping involves the addition of 7-methylguanosine to the 5' end of the mRNA; 3' polyadenylation, on the other hand, involves the addition of adenine bases to the 3' end to form a poly(A) tail. RNA splicing is the process by which introns (non-protein-coding regions) are removed from the pre-mRNA, resulting in the joining of exons (mostly coding regions) to form mature mRNA [5]. The inclusion of different combinations of exons normally leads to the production of multiple distinct functional isoforms from a single gene.

1.2 Epigenetic Regulation of Gene Expression

"Epigenetic" literally means "on the top" of genetic and epigenetic regulation is a mechanism that provides regulatory information to a genome without altering its primary DNA sequence. Two major epigenetic modifications, which tightly regulate gene activity, are the modification of the histone proteins associated with DNA (histone modifications) and the addition of a methyl group to the cytosine residues of DNA (DNA methylation) [6, 7]. The nucleosome, the basic unit of chromatin, contains two copies of each of the histone core proteins H2A, H2B, H3 and H4, which together form a histone octamer, and about 147 base pair (bp) of DNA wrapped around it [8]. Post-translational modification of these histones plays a key role in the regulation of gene activity and expression during development and differentiation. Histones can be modified by many different post-translationally added chemical groups like methylation, phosphorylation, acetylation, ubiquitination and sumoylation [8]. These modifications can influence chromatin structure and protein binding and thus, gene transcription.

Different histone modifications are broadly associated with activation or repression of gene expression. Acetylation and methylation of lysines at histone tails are the two most extensively studied modifications, with distinct distributions along both euchromatin and heterochromatin. For example, di- and tri-methylation of lysine 4 on histone 3 (H3K4me2 and H3K4me3, respectively) are generally associated with euchromatin and ongoing gene expression. On the other hand, tri-methylation of lysine 9 on histone 3 (H3K9me3) and tri-methylation of lysine 27 on histone 3 (H3K27me3) are mainly associated with heterochromatin and gene silencing [9]. Acetylation leads to a reduction of positive charges on the histone tails and loosens DNA-histone interactions leading to an open chromatin state, and is, therefore, generally found at actively transcribed promoters [10]. These histone modifications are established or removed by different families of enzymes. For example, histone methyltransferases (HMTs) catalyze the addition of methyl groups, and mainly contains the evolutionary conserved SET domain [11, 12]. On the other hand, histone demethylases (HDMs) can remove the methyl groups. The cross-talk between different histone modifications (or histone-modifying enzymes) can bring about distinct chromatin states, which, therefore, tightly regulates spatiotemporal gene expression [10, 13, 14].

The second extensively studied epigenetic modification is DNA methylation. Unlike histone modifications, which involves modification of the histone proteins, DNA methylation involves the addition of a methyl group at the 5' position of the cytosine ring to create a 5-methylcytosine (m5C). In mammalian cells, the majority of DNA methylation occurs on cytosines that precede a guanine nucleotide known as CpG dinucleotide sites [15, 16]. DNA sequences of several hundred to approximately two thousand base pairs with high frequency of CpG sites are commonly known as CpG islands [16]. Interestingly, in humans more than 50% of gene promoters harbor CpG islands [16]. DNA methylation is distributed throughout the genome and is linked to transcriptional silencing [17]. Normally, in mammals, many housekeeping or developmentally regulated genes have hypomethylated CpG islands in their promoters [16]. DNA methylation is catalyzed by DNA methyltransferases (DNMTs), including DNMT1, DNMT3a, and DNMT3b, which are responsible for its deposition and maintenance and are essential for normal development [16].

In addition to aforementioned histone modifications and DNA methylation, several other epigenetic mechanisms play an important role in regulating the gene expression (Figure 1.1). These include non-coding RNA (transcripts that are not translated), chromatin-remodeling complexes, and histone variants, for example [18, 19]. In recent

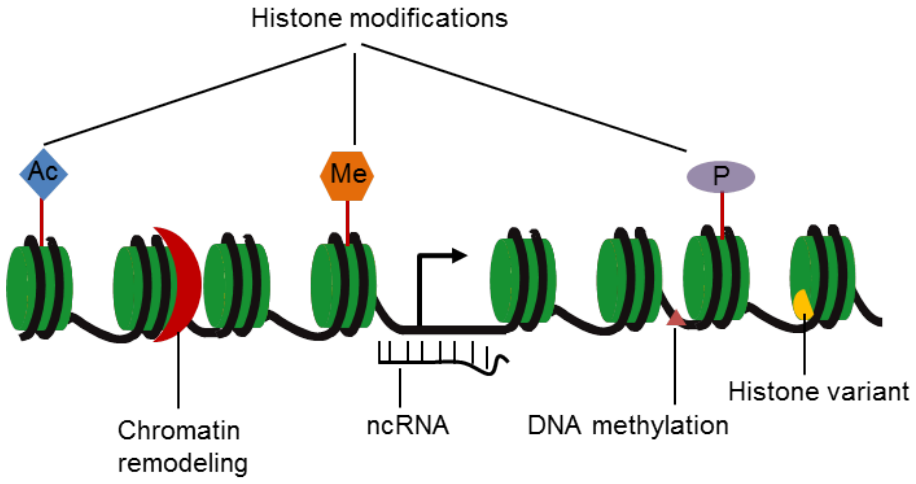


Figure 1.1: Different epigenetic mechanisms. The interplay of these different mechanisms play a critical role in gene regulation. Ac, acetylation; Me, methylation; P, phosphorylation; ncRNA, non-coding RNA.

years, it has become clearer that different epigenetic mechanisms work together to regulate gene expression. This epigenetic interplay is critical for normal development.

1.3 Human Heart and Skeletal Muscle Development

The heart is one of the most important organs and is, in fact, the first organ to form and function during development of vertebrates [20]. In humans, the heart starts beating in the fourth week after fertilization [21]. It is the pumping organ of the body, which circulates the blood. It has four chambers: the right atrium, left atrium, right ventricle, and left ventricle. The right atrium act as receiving chamber for deoxygenated blood from the body transported through the venae cavae (Figure 1.2). It pumps the deoxygenated blood to right ventricle and finally into the lungs. The left atrium receives the oxygenated blood from lungs and pump it to the left ventricle and finally back into the body through the aorta.

The development of the embryonic heart, also known as cardiogenesis, is a precisely controlled process, which includes a series of events including the formation of the heart tube, looping or bending events, chamber formation, septation and development of the valves [22–24]. The mature four-chambered heart consists of different cell types (Figure 1.2), including atrial cardiomyocytes, ventricular cardiomyocytes, smooth muscle cells, endothelial cells, epicardium cells, fibroblasts and pacemaker

cells [22, 24]. These different cell types are derived from multipotent cardiac progenitor cells, which can be divided into two main categories: the first population of cells to migrate to the heart-forming region, known as the primary heart field (or first heart field; FHF), and the second population of cells, which contributes progressively to the poles of the elongating heart tube, known as the secondary heart field (SHF) [24, 25]. The FHF gives rise to the left ventricle as well as the right and left atria, whereas the SHF contributes to a major part of the heart, forming the right ventricle, outflow tract, and the right and left atria. The formation of the vertebrate heart is controlled by crosstalk between multiple inter/intracellular signaling pathways and transcriptional regulatory networks in the FHF and SHF [24, 25].

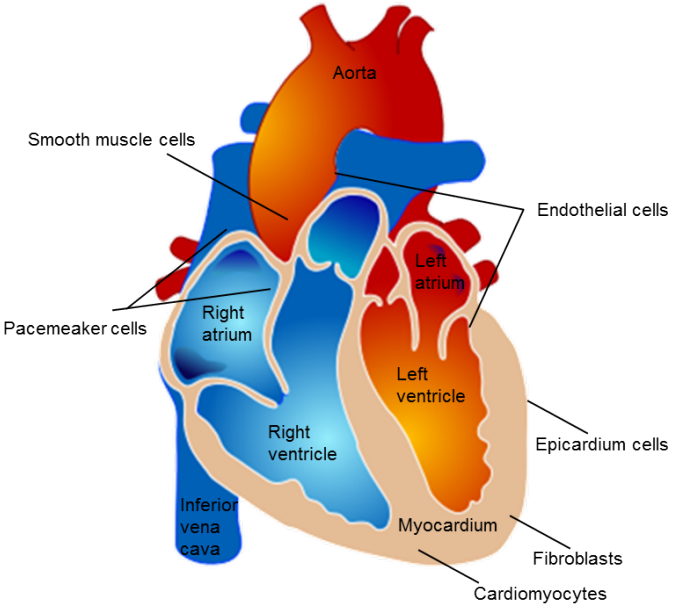


Figure 1.2: Schematic representation of the four-chambered human heart. Regions indicated in blue and red contain deoxygenated and oxygenated blood, respectively. The mature heart consists of different cell types, some of which are labelled. Figure taken from M. Ruiz [26] and modified.

A core set of crucial and evolutionarily conserved cardiac transcription factors, including the Nkx2-5, GATA family, Mef2 factors, Srf, Tbx-factors, Hand2 and Isl1, all play a central role in cardiac development [24, 27]. Numerous groups have generated various knockout mice, which are useful scientific tools to understand the roles of genes during development. For example, Gata4 homozygous knockout mice die at E8.5 with failure of ventral morphogenesis and heart tube formation [28, 29]. Targeted disruption of Nkx2-5 leads to abnormal heart morphogenesis and growth

retardation, with lethality at E9.5 [30]. Mice deficient in Mef2a die within the first postnatal week and exhibit myofibril fragmentation, pronounced dilation of the right ventricle, and impaired myocyte differentiation [31]. Cardiac-specific ablation of Srf in mice resulted in embryonic lethality due to cardiac insufficiency during chamber maturation [32]. Tbx5 deficiency in homozygous mice leads to the arrest of heart development at E9.5, resulting in lethality at E10.5 [33, 34]. Hearts of Isl1 homozygous knockout mice do not develop the outflow tract, right ventricle, and much of the atria [35]. Moreover, Vincentz et al. have shown a functional role for genetic Nkx2-5 and Mef2c interactions using Nkx2-5/Mef2c double null mice embryos that resulted in ventricular hypoplasia, a more severe cardiac phenotype than those associated with either single mutant [36].

Recently it has become clear that these core transcription factors interact with one another and provide cooperative regulation of individual target genes to control heart development [27]. A systems biology study published by the Sperling lab showed combinatorial regulation by Gata4, Mef2a, Nkx2-5, and Srf and demonstrated that they can partially compensate each other's function [37]. In addition to these cardiac transcription factors, the study integrated the mRNA profiles, microRNA profiles, and four activating histone modification marks (H3K9K14ac, H4K5K8K12K16ac, H3K4me2 and H3K4me3) in mouse HL-1 cardiomyocytes. They found several target genes associated with these TFs; specifically, 345 target genes for Gata4, 701 for Mef2a, 276 for Nkx2-5 and 1,150 for Srf. It was immediately evident that many of the target genes were shared by these factors; for example, Gata4 and Nkx2.5 shared 143 targets and Mef2a and Srf shared 320 target genes. Using RNA interference (RNAi) knockdown of one respective factor, they demonstrated that genes regulated by multiple transcription factors were significantly less likely to be differentially expressed. This suggests that these TFs can interact cooperatively to synergistically activate transcription of target genes. The misexpression of cardiac transcription factors or their cofactors can disrupt the gene regulatory networks, which may lead to cardiovascular disease.

Congenital heart diseases (CHDs) are the most common birth defect in humans, with an incidence of around 1% of all live births [38, 39]. Although, the exact etiology of CHD remains unclear, it is becoming clearer that genetic factors play an important role. Numerous studies have demonstrated the role of single gene defects in non-syndromic CHD. For example, mutations in GATA4, MYH6 or NKX2.5 have been associated with atrial septal defects [40–43]. In addition to genetic factors, environmental factors have been suggested to play a role. Almost five decades ago, Dr.

James Nora proposed that both genetic and environmental factors participate in the etiology of CHDs [44]. Despite the importance of environmental factors in CHD, most studies to date mainly focus on genetic factors because of the complexity of environmental factors, which makes it complicated to study and link them to the disease. Therefore, environmental factors often are viewed as noncontributory or secondary [45]. Nevertheless, factors like smoking during early pregnancy have been associated with an increased risk of CHD [45, 46]. Moreover, it has been suggested that environmental factors increase the risk of having a disease if genetic abnormalities are present [45]. Despite these links, it remains difficult to clearly associate different genetic and/or environmental factors with different types of CHD.

CHDs comprise a heterogeneous group of cardiac malformations and can be classified into three broad categories: cyanotic heart disease, left-sided obstruction defects, and septation defects [47]. The most common cyanotic form (blue skin color caused by a lack of oxygen) of CHD is Tetralogy of Fallot (TOF), which accounts for up to 10% of all heart malformations [48]. TOF (Figure 1.3) is characterized by four cardiac features: ventricular septal defect, overriding aorta, right ventricular outflow tract obstruction and right ventricular hypertrophy [49]. It is a well-recognized subfeature of syndromic disorders such as DiGeorge syndrome (22q11 deletion), Down syndrome, Holt-Oram syndrome and Williams-Beuren syndrome [50]. Deletions at the 22q11 locus account for up to 16% of TOF cases [51].

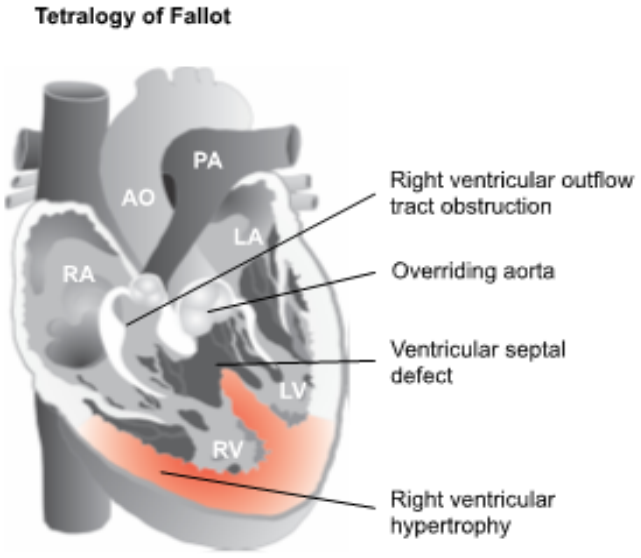


Figure 1.3: Heart with the 'Tetralogy of Fallot' phenotype, depicting the four clinical features. Figure taken from Grunert *et al.* [52].

It has been suggested that different genes with multiple mutations can result in a common phenotypic expression, which could be explained by the disruption of a common molecular network [52–54]. A recent study showed that isolated TOF is caused by a combination of deleterious private and rare mutations in neural crest (NC), apoptotic and sarcomeric genes [52]. These significantly affected genes (called TOF genes) coincide in an interaction network, which suggests that disturbances to a common network can lead to the phenotypic consequence of TOF. Moreover, recent studies demonstrated the role of copy number variations (CNVs) in the etiology of TOF cases [55–58]. Silversides et al. performed a large scale CNV analysis on 340 TOF cases and found that a significantly greater proportion of cases harbored large rare CNVs compared to controls [56]. They found CNVs in some of the interesting candidate genes; for example, copy number gain of GJA5 (Gap Junction Alpha-5 Protein) and a copy number loss of PLXNA2 (Plexin A2). In the future, it would be interesting to perform a multilevel study on a large cohort of isolated TOF cases to dissect the multifactorial etiology of the disease.

Unlike cardiac muscle, which is found in the heart, skeletal muscle is attached to the bone and is under voluntary control. Skeletal muscles play an important role in supporting and moving our body through contraction and relaxation. As in cardiac muscle, the sarcomere is the basic functional unit in skeletal muscle tissue that slide past each other when the muscle contract and relax [59]. The myofibril contains the repeating units of sarcomeres that are separated from one other by Z discs. Each individual sarcomere consists of many parallel actin (thin) and myosin (thick) filaments (Figure 1.4). These filaments slide in and out between each other during muscle contractions [59]. In 1954, two research teams independently describe the molecular basis of muscle contractions, which is known as the sliding filament theory [60, 61]. They observed that the length of the thick filaments of myosin remained relatively constant during contraction, whereas the thinner filaments made of actin change their length. Based on these observations, they proposed the model of muscle contraction (sliding filament theory), which states that the sliding of actin past myosin generates muscle tension. During the contraction, the length of the actin filament shortens, which results in a shortening of the sarcomere and thus, the muscle. Currently, this theory is the widely accepted model of muscle contraction.

The process of generating skeletal muscle is known as myogenesis. Progenitor cells originating in the somites give rise to skeletal muscle during embryogenesis [62]. Once these somites establish polarity they subsequently develop distinct dorsoventral compartments. The major dorsal part of the somite remains epithelial and turns into

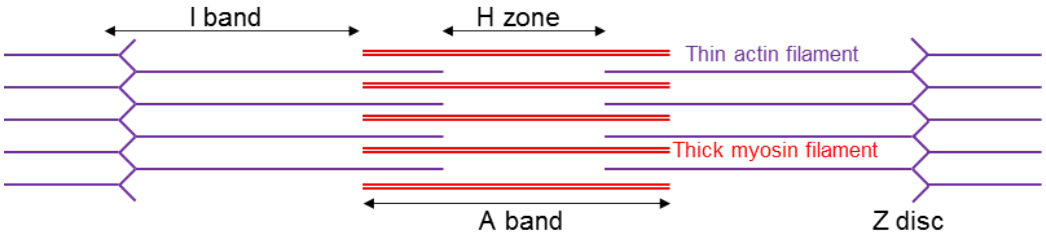


Figure 1.4: Basic structure of muscle fibre (sarcomere) sub-region. Actin (thin) and myosin (thick) filaments are shown in purple and red, respectively. During contraction, the length of H zone and I band shortens, whereas the length of the A band remains constant.

the dermomyotome [62]. The cells of the dermomyotome mature into the myotome, which leads to the development of all skeletal muscles of the body, with the exception of some head muscles [62]. This sequential development of the skeletal muscle involves the expression of different transcription factors at different stages. For example, the cells of the dermomyotome are marked by the expression of the paired box transcription factors Pax3 and Pax7; whereas, the cells of the myotome are marked by high expression of the basic helix-loop-helix myogenic regulatory factors (MRFs) Myf5 and MyoD [62–67]. MyoD and Myf5 are both considered to be markers of terminal specification and early differentiation, whereas another two MRFs, myogenin and Myf6, are considered to be markers of late differentiation [62, 68]. These four highly conserved transcription factors are together known as the myogenic regulatory factors and are collectively expressed in the skeletal muscle lineage [62, 69].

Almost three decades ago, MyoD was demonstrated to induce the conversion of fibroblasts into muscle cells [70], and was subsequently considered to be a master regulator of myogenesis. MyoD is expressed at the time of myogenic specification and binds to DNA via a consensus E-box motif (CANNTG). Upon the induction of differentiation, MyoD forms heterodimers with members of the E-protein family, with an increased affinity at many regulatory elements of skeletal muscle-specific genes [71, 72]. In undifferentiated myoblasts, MyoD and Baf60C, a subunit of the ATPase-containing SWI/SNF remodeling complex, form a complex on MyoD target promoters and mark genes prior to the activation of transcription, which play a role in myogenic differentiation [73]. During myogenesis, the binding of MyoD is primarily associated with gene activation [74], but its repressive function in myogenesis has also been shown on single genes [75–77]. In addition to promoters, genome-wide analysis has indicated MyoD binding events in intergenic regions in myoblasts and myotubes

[78, 79]. Moreover, the presence of MyoD is also highly associated with muscle-related enhancers [80].

In the past decade, number of different research teams has performed genome-wide epigenetic analysis during myogenic differentiation. It was shown that the overall content of histone methylations such as H3K4me2, H3K4me3, H3K36me3 and H3K27me3 remains stable during myogenic differentiation [81]. The repressive histone mark H3K27me3 was found to be widely distributed throughout the genome and regulates myogenic differentiation via silencing of muscle-specific and cell cycle genes [81–83]. However, histone 3 acetylations like H3K9ac and H3K18ac are reduced in a differentiation-dependent manner [81]. Interestingly, the regions of increased histone 4 acetylation have been associated with the genome-wide binding of the transcription factor MyoD [79].

1.4 High-Throughput Sequencing

In 1977, Frederick Sanger and colleagues published a groundbreaking method, known as Sanger sequencing, for determining nucleotide sequences within a DNA molecule [84]. This method requires modified di-deoxynucleotidetriphosphates (ddNTPs), which terminate elongation of the DNA strand elongation; thus, the method is also referred to as the chain termination sequencing method. For more than two decades, it remained the most widely used sequencing method and it underwent many technological improvements. These advances led to the development of the semi-automated Sanger method (also known as first-generation technology), which had higher throughput, enabling the completion of the first human genome sequence [85, 86]. The Human Genome Project (HGP), which began in 1990, aimed to determine the highly accurate sequence of the vast majority of the euchromatic portion of the human genome and was completed in 2004 at the cost of about US$3 billion [86–88]. As the HGP took 14 years to complete, it soon became clear that faster and cheaper technologies, with higher throughput, needed to be developed. Therefore, in the same year (2004) the National Human Genome Research Institute (NHGRI) aimed to reduce the cost of human genome sequencing to US$1000 in 10 years [86, 89]. This triggered the development and commercialization of second-generation or, more commonly, next-generation sequencing (NGS) technologies. Indeed, this goal was achieved through collective efforts, and the cost of sequencing has been reduced tremendously using NGS [86].

Over the past decade, the NGS market has been dominated by three major platforms or companies: Roche 454, Illumina, and Sequencing by Oligonucleotide Ligation Detection (SOLiD). All of these platforms depend on the preparation of NGS libraries (fragments of DNA or RNA) in a cell free system; subsequently, thousands-to-many-millions of sequencing reactions are produced and directly detected in parallel [86]. There are different sequencing mechanisms used by these platforms, all of which have advantages and disadvantages. For example, Roche 454, released in 2005, uses a pyrosequencing method based on the detection of light emitted by the release of pyrophosphate when a nucleotide is incorporated. This platform takes relatively less time to run and gives long reads of up to 1 kb, which makes it easier to map the reads to the reference genome [86, 90, 91]. On the other hand, it has relatively low throughput and a high error rates in homopolymer repeats [86, 92]. In 2006, Illumina/Solexa released its sequencing platform based on sequencing-by-synthesis method in which four differently labelled, reversible terminator-bound dNTPs are used. When nucleotides are incorporated, they are identified by color and, subsequently, the terminator is removed. As compared to other platforms, Illumina offers the highest throughput and the lowest per-base cost [86, 93] but can take long time (27 hrs to 11 days) to run. One year later, in 2007, the third technology, SOLiD, was released by Applied Biosystems (now Life Technologies). This platform uses fluorescently labeled octamers and color detection; each base is read twice and therefore claimed to have high accuracy (99.94%) [86, 93]. Despite this advantage, it has a relatively higher cost than the others. Among these platforms, Illumina remains the leader in the NGS industry because of its high-throughput and lower cost. It is worth noting that Roche decided to shut down 454 by mid-2016 [86] but a similar technology was released by Ion Torrent (now Life Technologies) based on the detection of proton instead of pyrophosphate, which does not require optical scanning. A comparison of these platforms, along with an automated Sanger sequencing machine, is depicted in the Table 1.1.

With the advancement of sequencing methods, from first-generation (automated Sanger sequencing) to second-generation, the term next-generation sequencing has become more common and widely used by researchers and companies [85, 92, 94–97], instead of high-throughput sequencing. Considering the pace of technological development, this term itself may soon be outdated. One example is the PacBio RS instrument released by Pacific Biosciences in 2010, which allows the detection of single molecule and is therefore considered to be the third-generation sequencing technology [86, 98]. In contrast to previous sequencing technologies, which require

	Illumina	Life Technologies	Roche/454	Life Technologies
Platform	HiSeq 2000/2500	SOLiD 5500/5500xl Wildfire	GS FLX+	Sanger 3730xl
Sequencing mechanism	Sequencing-by-synthesis	Ligation and two-base coding	Pyrosequencing	Dideoxy chain termination
Read length (bp)	50SE, 50PE, 100PE, 150PE	50SE, 75SE, 50PE	Up to 1,000 bp	400 to 900 bp
Reads	3 B/ 6 B single reads per run (2 flow cells)	1.6 B/ 3.2 B single reads per run	1 M single reads per run	_____
Run time	27 hrs to 11 days	10 days	23 hrs	20 mins to 3 hrs
Advantage	High throughput	Accuracy	Read length, fast	High quality, long read length
Disadvantage	Short read assembly	Short read assembly	Error rate with polybase more than 6, high cost, low throughput	High cost, low throughput

Table 1.1: Overview of the different sequencing platforms. The information is collected from Liu *et al.* and Dorn *et al.* [90, 91]. B, billion; M, million; bp, base pairs; hrs, hours; mins, minutes; SE, single-end reads; PE, paired-end reads.

multiple identical copies of a DNA molecule, researchers from Pacific Biosciences demonstrated the Single Molecule, Real-Time (SMRT) technology that can sequence a single molecule of DNA. This technology utilizes the zero-mode waveguide (ZMW), which reduces the volume of observation to the point where it is sufficient enough to observe only a single nucleotide of DNA being incorporated by DNA polymerase [86, 98]. A major advantage of this method is the extremely long reads of 4-40 kb, which could potentially help to improve the existing draft genomes [86]. On the other hand, it has high overall error rates and is relatively expensive [86].

Despite the limitations, high-throughput sequencing methods are extremely useful for addressing a large range of biological questions. With rapid advances in the technology, the applications of these methods seem almost endless. Using this technology, it is now possible to sequence an entire genome in less than one day. One of the most widely used applications is the analysis of the genome to identify sequence variations in genes and regulatory elements. These mutations can be of different types and can affect anywhere from a single nucleotide (base pair) to a large segment of a chromosome; for example, change of a single nucleotide (also known as single nucleotide variations or SNVs), deletion of a piece of DNA (copy number loss), duplication of a

piece of DNA (copy number gain), or insertions/deletions (indels). Most often, the main goal is to identify these variations and what role, if any, they play in the disease. Some of the other main applications of high-throughput sequencing are discussed in the next chapter.

The field of Bioinformatics includes development and improvement of methods for storing, retrieving, organizing and analyzing biological data generated using high-throughput sequencing (HTS) technologies. The major focus is to generate useful biological knowledge in an efficient manner. Data storage and accuracy of data analysis are the major challenges in the field of Bioinformatics. From the very first step of the data analysis meaning alignment of the reads, it is challenging to develop the software which can be used for multiple projects using different HTS technologies. Moreover, HTS technology is rapidly evolving that requires constant improvement of existing methods. Furthermore, the gap between the mass generation of data and the ability to analyze this data is growing. Therefore more efforts are required for the comprehensive analysis of the data to answer precise biological questions.

1.5 Aims of the Thesis

High-throughput sequencing (HTS) technology is rapidly evolving and revolutionizing research in the life sciences. Due to its low cost and high throughput, HTS is used commonly by various laboratories to answer different biological questions. With the advancement of sequencing platforms, there is an increase in demand on statistical methods and computational approaches for analyzing HTS data, namely targeted DNA resequencing data, RNA-seq data and ChIP-seq data. The goal of this thesis is to establish computational approaches for analyzing HTS read count data, aimed at answering concise biological questions.

Congenital heart diseases (CHD) are the most common birth defect in human, with an incidence of around 1% of all live births. The most common cyanotic form of CHD is Tetralogy of Fallot (TOF), which accounts for up to 10% of all heart malformations. Previous studies have demonstrated the role of copy number variations (CNVs) in the etiology of TOF. The first study in this thesis (Chapter 4) aimed to identify copy number alterations in a small cohort of non-syndromic TOF patients based on targeted resequencing data. Detecting CNVs from targeted resequencing data is difficult due to nonuniform read-depth between captured regions. Moreover, there was no tool available to detect personalised CNVs from small cohort of patients without using controls. Therefore, a novel copy number variation calling method

was developed to identify individual disease-relevant copy number variations (CNVs) using exome or targeted resequencing data of small sets of samples.

Myogenic differentiation is an essential process of muscle development and depends on the spatiotemporal regulation of gene expression patterns. The interplay of transcription factors and the chromatin changes is an important attribute to govern gene expression. Using ChIP-seq and RNA-seq data, systematic analysis was performed (Chapter 5) to investigate a stable enrichment pattern of the histone marks H3K4me2 and H3K4me3 in combination with muscle tissue-specific transcription factor MyoD during myogenic differentiation. The final study in this thesis (Chapter 6) aimed to develop a pipeline to identify differential exon usage from RNA-seq data, with the intention of identifying the exons either excluded or included. In previous studies, Sperling lab identified a chromatin remodeling factor Dpf3, the expression of which was significantly up-regulated in the right ventricle of TOF patients [99, 100]. Therefore, the final study in this thesis (Chapter 6) dissected the role of Dpf3 in splicing.

Chapter 2

High-Throughput Sequencing Methods and Computational Analysis

2.1 Methods

The advent of the high-throughput sequencing (HTS) technology has greatly accelerated research in life sciences. Due to its low cost and high throughput, HTS is commonly used by many laboratories to answer biological questions. Moreover, it is now possible to sequence the entire human genome in less than one day. Besides whole genome sequencing, HTS has other applications like the identification of genome-wide protein-DNA interactions and quantification of the gene expression. In general, HTS is used to determine the sequence of millions of DNA fragments in parallel, and these fragments can be generated using various methods. All used methods and computational analysis are described in this chapter.

2.1.1 Targeted Resequencing of Genomic DNA: DNA-seq

Although the cost of sequencing the entire genome has decreased tremendously over the past ten years, whole genome sequencing is not the only method used. Researchers often prefer to select specific regions of interest and enrich these regions for sequencing. For example, in whole exome sequencing (WES), the protein-coding regions are targeted and sequenced. In humans, about 1% of the genome is coding (exome), and sequencing only these regions makes it more cost-effective [101, 102]. Moreover, it has been suggested that the exome might contain ∼85% of known disease-causing variants, making this method more attractive for detecting causal variants [101, 102].

Interestingly, a market research survey carried out in 2013 by Oxford Gene Technology showed that the top preferred and used method is targeted resequencing, which includes WES and targeted panel sequencing [103]. In the latter, instead of selecting the whole exome, a specific group of interesting genes are enriched and sequenced. There are several approaches available for target enrichment such as PCR-based amplification, molecular inversion probe-based amplification (MIPs), and hybridization-based sequence capture (array-based and in-solution) [104–107]. One of the most widely used approaches is array-based hybridization capture method [108–110], which is also used in this study and is described below.

Targeted capture consists of the library preparation step, where fragments of genomic DNA are ligated with the adapters. This library is hybridized to the sequence capture array (Figure 2.1). The capture array contains the immobilized probes, which are single stranded DNA molecules attached to a solid surface. Different sequence capture arrays can have different number and/or length of probes. For example, Roche NimbleGen capture array can have 385,000 isothermal probes (385K array) with a total capture size of up to 5 Mb or 2.1 million probes with a total capture size of up to 34 Mb [104]. After the hybridization of DNA fragments with the probes, unbound fragments are removed by washing and the enriched fragments are eluted [104]. Amplification of these enriched fragments is performed using ligation-mediated polymerase chain reaction (LM-PCR). Before sequencing of these amplified enriched fragments, a quality control step is carried out using quantitative PCR (qPCR) at control loci. Initially, the method was designed to be used with the Roche 454 sequencer but with modified and optimized protocols, the Illumina platform can also be used [104].

Features like high sequencing accuracy, low cost, coverage depth, experimental focus, and sample number make targeted resequencing more and more popular. Moreover, the use of this method makes the downstream computational analysis more feasible and gives a less complex outcome that is functionally interpretable. Targeted panel sequencing, in which only a panel of interesting genes are targeted and coupled with high-throughput sequencing, is revolutionizing the clinical research. One interesting example is “TruSight Cardio Sequencing Kit” by Illumina, whic is useful for the identification of variants in 174 genes at a cost of approximately $1 US per gene. In the future, these customized gene panels can be used regularly as a diagnostic tools [111].

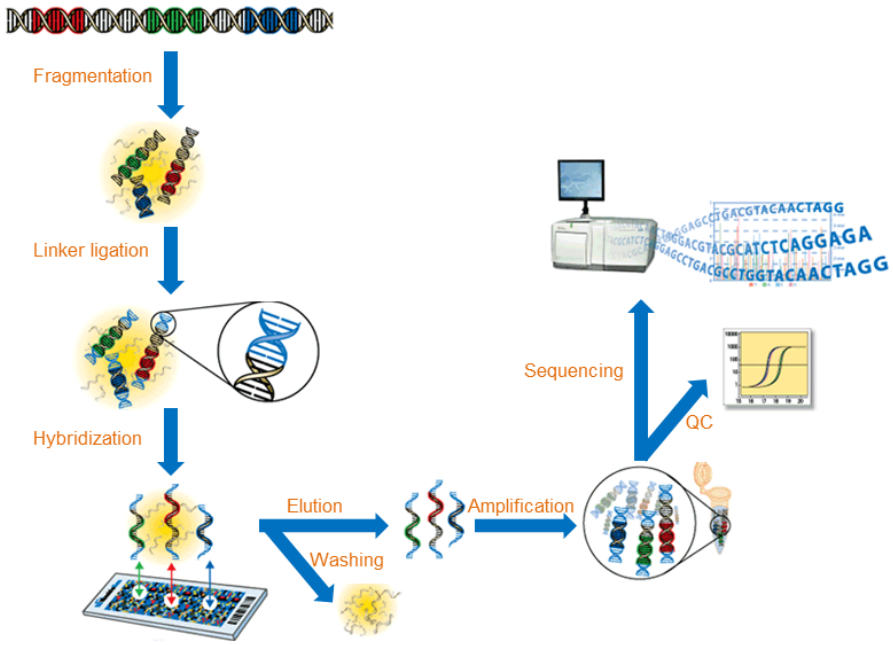


Figure 2.1: Roche NimbleGen capture array technology for the enrichment of genomic target regions. Figure taken from NimbleGen [112] and modified.

2.1.2 Quantification of Gene Expression: RNA-seq

The transcriptome is the total set of transcripts in a cell, including mRNAs, long and short non-coding RNAs and small RNAs like microRNAs [113]. Using high-throughput sequencing, we can deduce and quantify the transcriptome of a population of cells and even compare it across multiple samples [113, 114]. In the market research survey carried out by Oxford Gene Technology [103], mentioned earlier, it was shown that RNA sequencing (RNA-seq) is the second most used method (after targeted resequencing) among researchers. RNA-seq involves direct sequencing of complementary DNA (cDNA). In general, a library of cDNA fragments is prepared from a population of RNA (total or fractionated, such as poly(A)+). More often, polyadenylated RNAs (poly(A)+) are captured and converted to stable cDNA fragments, which are then sheared, selected and amplified with adaptors attached to one or both ends [113, 114]. Finally, this library is sequenced from one end (single-end sequencing) or both ends (pair-end sequencing), using high-throughput sequencing technology to obtain short sequences or reads [113, 114]. This protocol is also known as PolyA-seq or mRNA-seq.

Although the standard protocol of RNA-seq library generation is commonly used, it loses the strand of origin information for each transcript. It has been suggested that the transcription of the DNA sense strand produces antisense transcripts, which often results in the production of non-coding RNAs (ncRNAs) [115]. These ncRNAs are complimentary to their associated sense transcripts, and it is therefore crucial to keep the information of the strand from which transcript originated. In order to overcome this issue, one can use a modified standard RNA-seq protocol, which is commonly refered to as strand-specific RNA-seq [115, 116]. One of the strand-specific or directional RNA-seq library preparation protocols is illustrated in Figure 2.2.

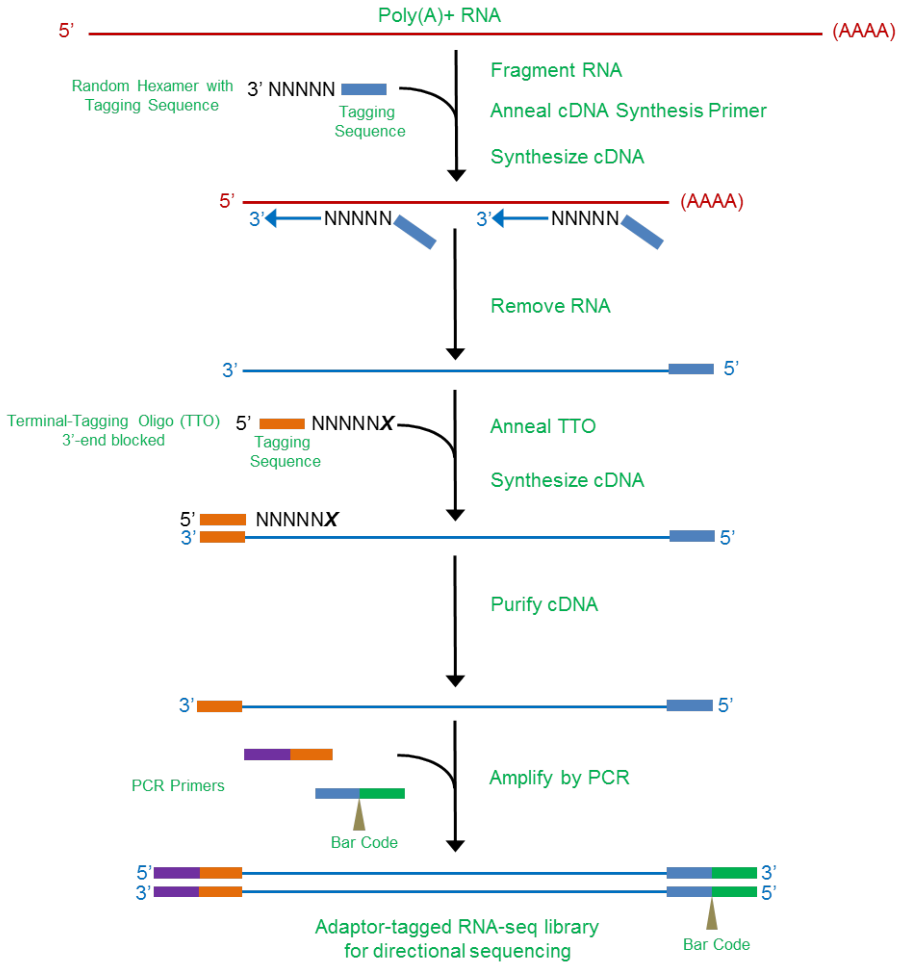


Figure 2.2: Illustration of directional RNA-seq library using "ScriptSeq RNA-seq library preparation kit" from Illumina. Figure drawn on the basis of [117].

2.1.3 Genome-wide Identification of Protein-DNA Interactions: ChIP-seq

Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing allows genome-wide identification of protein-DNA interactions such as transcription factor binding, transcriptional co-factor binding, RNA polymerase binding and chemical modification of histone proteins [118–122]. In a typical ChIP experiment (Figure 2.3), firstly the cells are treated with formaldehyde to cross-link the DNA-binding proteins to DNA [123]. Afterwards, the cross-linked strands are exposed to sonication, which fragments or shears the DNA using high-frequency sound waves. These DNA fragments, which are bound by different proteins, are immunoprecipitated using an antibody that recognizes a specific transcription factor or histone modification [123]. This results in a collection of all DNA fragments bound by the protein of interest, with non-specific fragments and proteins washed away. In the next step, the collected fragments are reverse cross-linked to remove the bound proteins (ChIP sample). More often, for comparison, an additional sample, known as "Input", is prepared in parallel, which is not immunoprecipitated [123]. Finally, using these samples, genome-wide analysis can be performed by high-throughput sequencing (ChIP-Seq). Firstly, the sequencing library is generated using the immunoprecipitated sample and input sample, which can be analyzed using high-throughput sequencing machines. The identification of protein binding sites (peaks) is carried out by comparing the number of sequenced reads generated from the immunoprecipitated sample and input sample [123].

In the past, ChIP followed by microarray hybridization (ChIP-chip) was commonly used to perform genome-wide mapping of protein-DNA interactions. For ChIP-chip, the immunoprecipitated sample and input DNA, are labeled with fluorescent dyes and hybridized to microarrays [123]. The identification of protein binding sites is carried out by comparing the intensity of signal of the ChIP samples to the signal of the input sample at each probe on the microarray [123]. With the advancement of sequencing technology, ChIP-seq has emerged as an attractive alternative to ChIP-chip due to its higher resolution and reduced noise [125]. Moreover, for ChIP-seq, no prior knowledge of the target DNA binding sites is required; however, despite these positives, ChIP-seq suffers from high cost and requires a large amount of starting material as compared to ChIP-chip [126]. Large scale projects like ENCODE (Encyclopedia of DNA Elements) and modENCODE (Model Organism ENCyclopedia Of DNA Elements) have used the ChIP-seq as the primary method and provide a set of standards and guidelines for performing ChIP-seq [127].

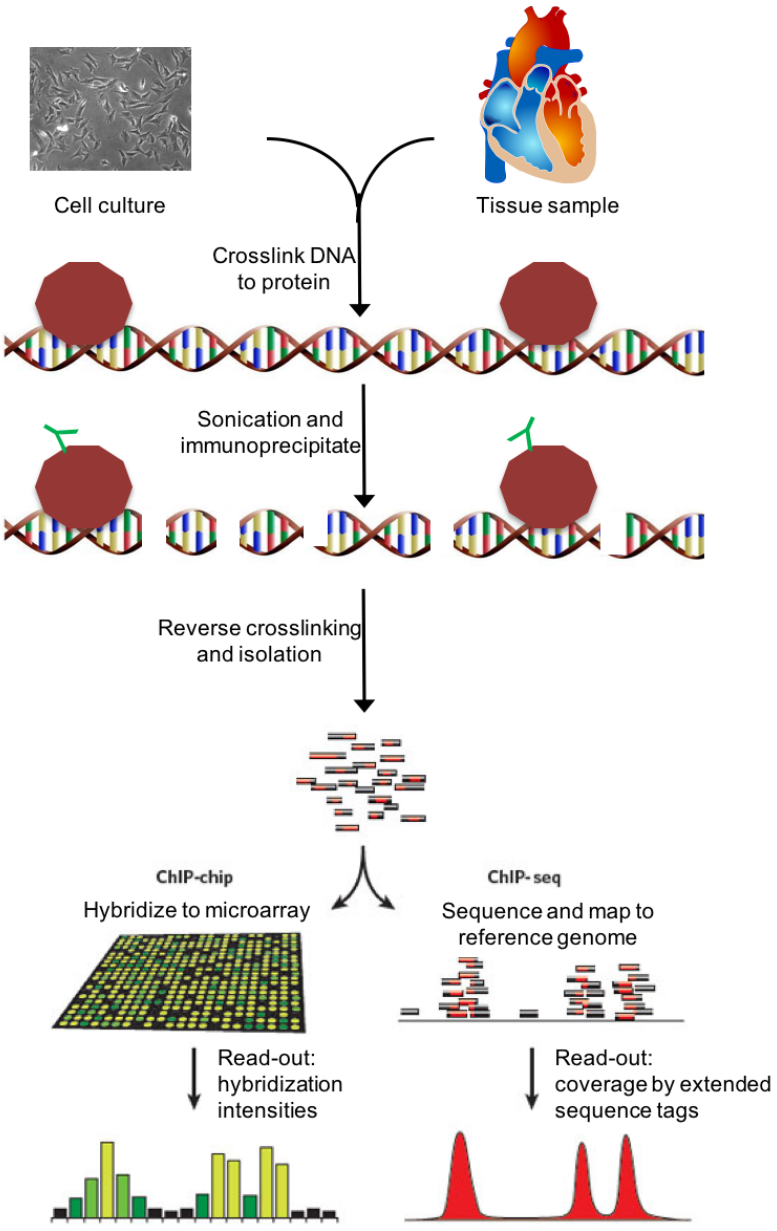


Figure 2.3: Schematic representation of a chromatin immunoprecipitation (ChIP) experiment followed by microarray hybridization (ChIP-chip) or high-throughput sequencing (ChIP-seq). Figure drawn on the basis of Visel *et al.* [124] and modified.

2.2 Computational Analysis of High-Throughput Sequencing Data

2.2.1 Alignment to a Reference Sequence

High-throughput sequencing (HTS) technology is rapidly evolving and revolutionizing research in the life sciences. HTS generates millions of short sequences (reads) that need extensive computational analysis to fetch out the information from the data. Usually, the first step in high-throughput sequencing data analysis is the alignment (mapping) of the generated reads to a reference sequence. The aim of the mapping is to find the location of reads from where they originated in the reference sequence. The read mapping problem can be generally stated as follows: given a set of read sequences Q, a reference sequence G and a possible set of constraints and a distance threshold k, find all substrings m of G that respect the constraints and that are within a distance k to a sequence q in Q. The occurrences m in G are called *matches* [128, 129]. The mapping process is complicated by several factors, including sequencing errors, genetic variations, short read length, multi-mapped reads, and the huge amount of reads to be mapped [129, 130]. Therefore, during the past decade, numerous software tools have been developed to accomplish this task efficiently (e.g. for DNA mappers: SOAP [131], MAQ [132], Bowtie [133], BWA [134], SHRiMP [135], RazerS [128], mrFAST [136]; and for RNA mappers: TopHat [137], SpliceMap [138], MapSplice [139], SOAPsplice [140], STAR [141]). The most widely used DNA mappers are Bowtie (6,117 citations) and BWA (6,009 citations), and the most popular RNA mapper is TopHat (3,244 citations). The following is a brief description of these software tools.

Both Bowtie and BWA, are full-text minute-space (FM) index based aligners. They employ a Burrows-Wheeler index based on the FM-index. FM-index is a compressed, yet searchable suffix array-like structure [142] from the Burrows-Wheeler transform of the reference genome [143]. Burrows-Wheeler transformation (BWT) is a reversible permutation of the characters in a text in such a way that characters from repeated substrings would be clustered together [133]. For example, the BWT of a text T or BWT (T), is constructed by appending the character $\$$ to T, where $\$$ is not in T and is lexicographically less than all characters in T [133]. The Burrows-Wheeler matrix of T is constructed as the matrix whose rows comprise all cyclic rotations of $T\$$ and transform is done by sorting all rotations of the text into lexicographic order. The sequence of characters in the last column of the Burrows-Wheeler matrix is the BWT(T) (Figure 2.4) [133].

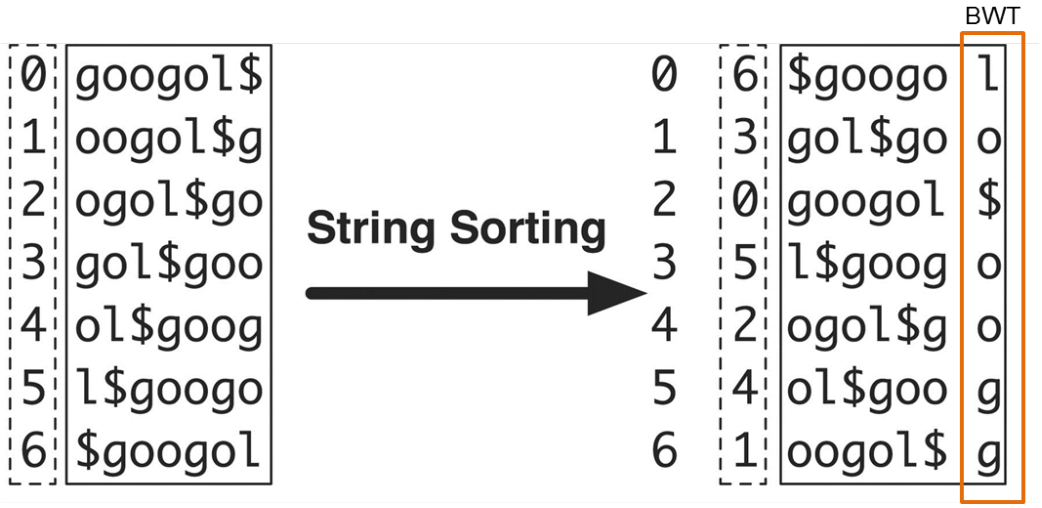


Figure 2.4: The Burrows-Wheeler matrix and transformation for "googol". Firstly, the character *$* is appended to the string and then cyclic rotations are carried out (left). Next, all the rotations are sorted into lexicographic order and BWT is the sequence of the characters in the last column (marked in orange). Figure taken from Li *et al.* [134] and modified.

The common method for searching in an FM index is the exact-matching algorithm, which search for only exact matches. Due to sequencing errors and genetic variations, we may not find exact matches for all the reads. Therefore, BWA and Bowtie uses the modified matching algorithms i.e. backtracking algorithm for BWA and quality-aware backtracking algorithm for Bowtie. BWA searches for matches between the read and the corresponding genomic position within a certain defined distance whereas Bowtie uses a quality threshold [133, 134, 144, 145]. It is important to note that these software tools are being updated regularly and new versions could have modified algorithms. For example, Burrows-Wheeler Aligner's Smith-Waterman Alignment (BWA-SW) can align long reads up to 1 Mb against a reference genome [146] and Bowtie 2 was mainly designed to map reads longer than 50 bps and supports gapped alignments [147].

TopHat (or the latest version TopHat2) is a fast splice junction mapper for RNA-seq reads [137, 148]. In general, TopHat is a pipeline to map RNA-seq reads to transcriptome and/or genome using Bowtie (or Bowtie 2) and then analyzes the mapping results to identify splice junctions between exons. The steps involved in the TopHat2 pipeline are depicted in Figure 2.5.

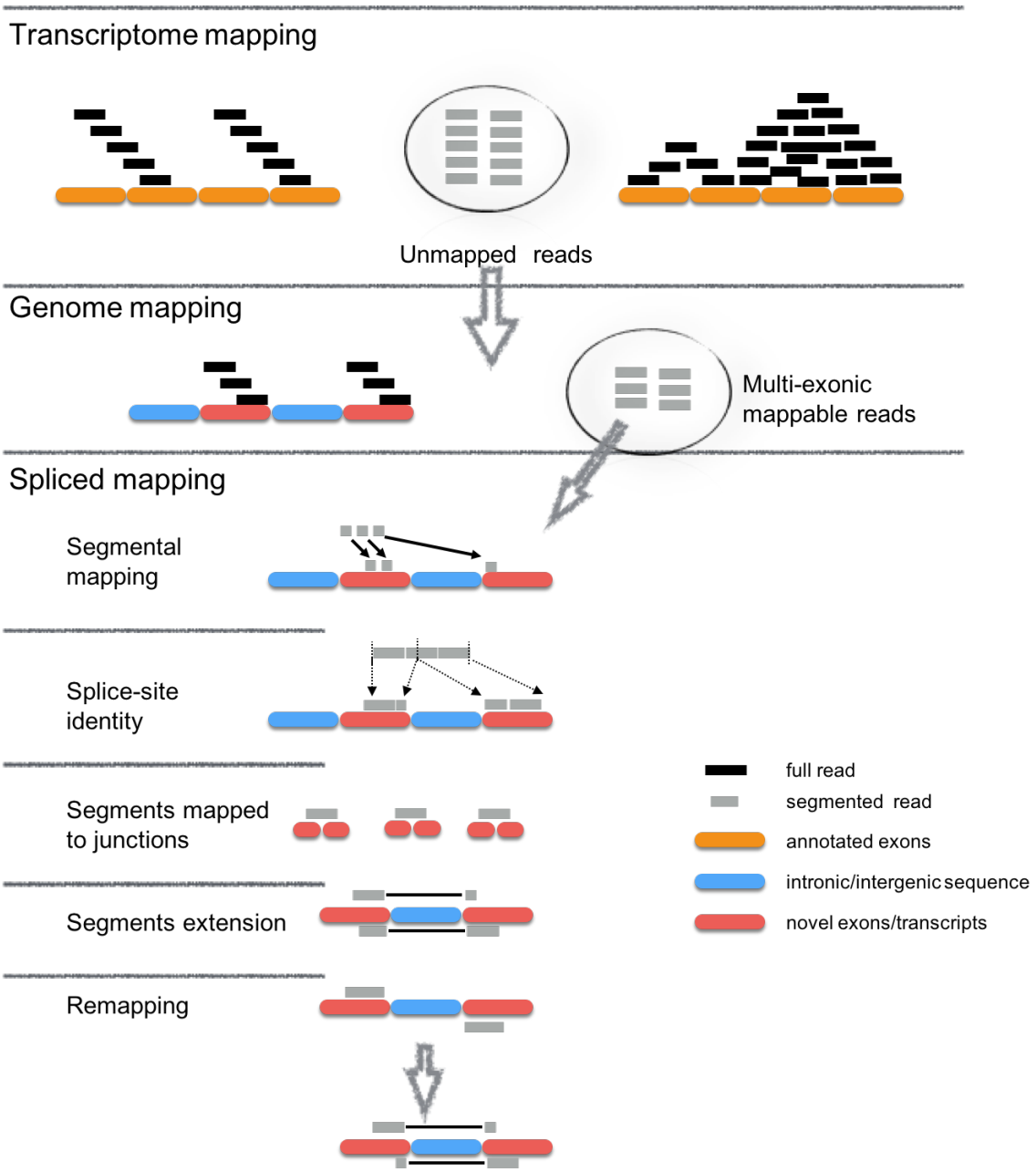


Figure 2.5: Illustration of steps involved in mapping RNA-seq reads using TopHat2. TopHat2 pipeline uses Bowtie (or Bowtie 2) to align the reads to reference transcriptome and unmapped reads are then aligned to the reference genome. Figure drawn on the basis of Kim *et al.* [148] and modified.

2.2.2 Analysis of DNA-seq Data

High-throughput sequencing technologies open a way to analyze millions of DNA sequences in parallel to detect genetic mutations. In general, after mapping the DNA reads to a reference sequence, we can investigate the differences between the sequenced reads and the reference sequence. Genetic mutations can alter the single base (Single Nucleotide Variations), can delete or duplicate a DNA sequence (Copy Number Variations) or can invert the DNA sequence (Inversion) or can insert or delete a small sequence of 2 to 50bp (indels). These mutations may have no phenotypic effect, account for adaptive traits or can cause disease. In this study, targeted resequencing data from patients with Tetralogy of Fallot was analyzed to detect Copy Number Variations (CNVs). CNVs are regions of a genome present in varying number in reference to another genome or population. In the last years, several computational strategies have been developed for detecting CNVs from DNA-seq data. For exome sequencing or targeted resequencing, the read depth or depth of coverage approach is widely used and described below.

2.2.2.1 Detection of Copy Number Variations

Copy Number Variations (CNVs) have been associated with a number of human diseases such as Crohn's disease, intellectual disability, cancer and congenital heart disease [149–153]. Microarray-based comparative genomic hybridization (array-CGH) allows analysis of the genome to identify CNVs without using high-throughput sequencing technologies (Figure 2.6). This method compares a patient DNA with a reference DNA (normal control) which are differentially labelled using fluorescent dyes. Next, the patient DNA and the reference DNA are hybridised on a microarray containing the oligo (oligonucleotide) probes [154]. Each probe represents a specific locus in the genome. The DNAs will bind to probes with complementary sequence. After hybridisation, the fluorescence of each dye for each probe is measured and the relative intensity between the two fluorescent dyes is calculated for each probe. If the intensities for the two dyes are equal for a given probe, it is considered as the normal copy number. An altered intensity for the patient DNA represents a loss or a gain of the patient DNA at that specific genomic region. Apart from array-CGH, DNA-seq data is widely used for detecting CNVs which is described below.

With the advancement in high-throughput sequencing technologies, several computational strategies have been developed for detecting CNVs from high-throughput

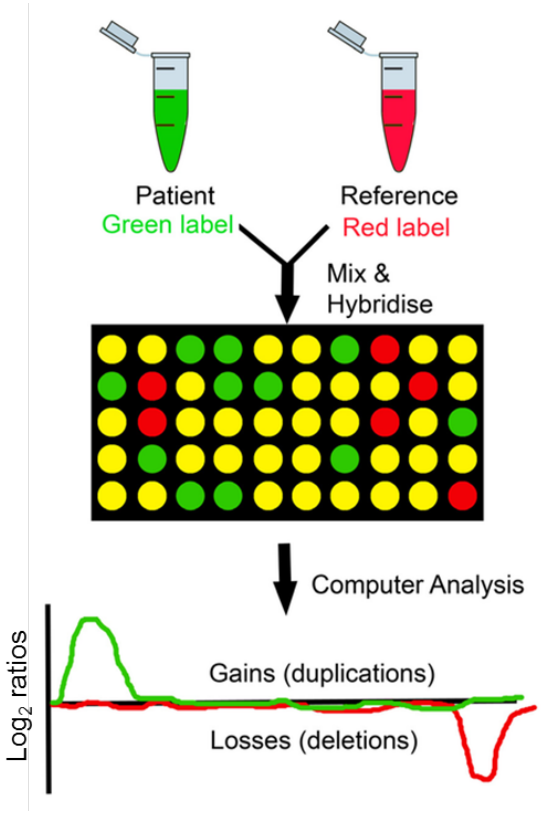


Figure 2.6: Microarray-based comparative genomic hybridization. Figure taken from Karampetsou *et al.* [154] and modified.

sequencing data. One of the widely used method is the read-depth approach (Figure 2.7) which is able to detect very large gains (duplications) or losses (deletions) and works on single-end as well as paired-end data [155–157]. It assumes that the mapped reads are randomly distributed across the reference genome or targeted regions. Based on this assumption, the read-depth approach analyses differences from the expected read distribution to detect duplications (higher read depth) and deletions (lower read depth) [157]. In simple terms, for example, we can divide the whole reference sequence into 10 windows and count the number of reads in each window. From this, we can calculate the average number of reads, which is an expected read count. Then we can compare the expected read count with the observed read count in each window and the windows with a higher and lower read depth are potentially duplicated and deleted regions, respectively. Depending on the sequencing method and technology used, the differences in the read depth might represent technical noise. Therefore, some normalization steps are necessary before CNV calling, which are described below.

Using the read-depth approach, several tools have been developed to identify

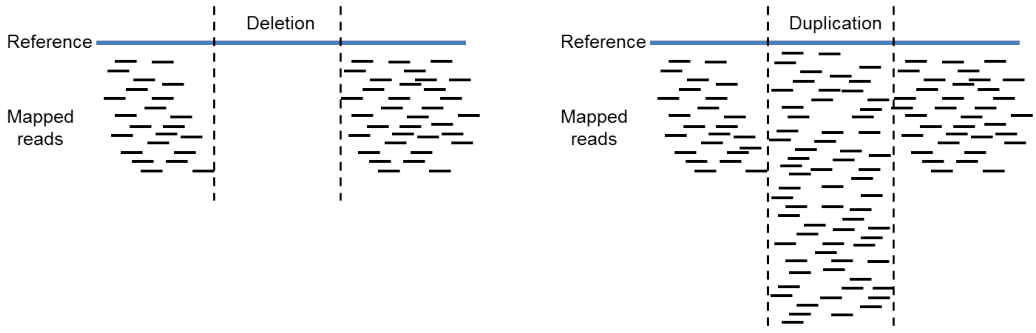


Figure 2.7: Read-depth approach for detecting copy number variations from DNA-seq data. The reference sequence (in blue) has been divided into three windows and mapped reads in each window has been shown below (small black lines). Deleted region (on left) has no reads or can have very few reads, whereas duplicated region (on right) has a higher number of mapped reads.

CNVs from whole genome and/or exome sequencing data [136, 158–167]. As mentioned earlier, differences in read depth in a genomic region can be related to several biases such as local GC-content, as well as sequence complexity and sequence repetitiveness in the genome [168]. For whole genome sequencing (WGS) it has been shown that normalizing read depth against GC-content can be sufficient to predict CNVs accurately [136, 149, 156, 164, 169]. One of the tool, which uses the read-depth approach along with the normalization against GC-content is mrCaNaVaR (micro-read Copy Number Variant Regions) [136]. This tool works with the WGS data and can predict the absolute copy number for all the genomic intervals. Briefly, as a first step, this tool divides the reference sequence into windows and calculate GC-content for each window. Next, it calculates the read depth (read count) for each window using a mapped file. Reads spanning the border of two windows are assigned to the left window that contains the 5'-end of the read. Finally, it performs the GC correction of the read depth values and predict the absolute copy number over the windows [170]. GC correction is performed using the "LOESS model", which depends on the relation between the read count and GC content of the windows. For a given window, GC is the fraction of G and C bases in that window according to the reference genome. The GC bias curve is determined by loess regression of count by GC-content of windows [168].

Detecting CNVs from targeted resequencing data is difficult due to nonuniform read-depth between captured regions. The two well known exome sequencing based CNV detection tools, which are also used in this study, are CoNIFER (Copy Number Inference From Exome Reads) and ExomeDepth [161, 167]. CoNIFER uses singular

value decomposition (SVD) to eliminate biases in exome data and detect CNVs and genotype the copy-number of duplicated genes from exome sequencing data [167]. SVD is a method for data reduction and used for identifying the dimensions along which data points exhibit the most variation. Therefore, it is possible to find the best approximation of the original data points using fewer dimensions. Firstly, CoNIFER calculates the RPKM (reads per thousand bases per million reads sequenced) for each targeted region. Next, it transforms these RPKM values into the standardized z-scores, which they termed as ZRPKM values and SVD normalization is performed on these ZRPKM values. Using CoNIFER, the exome sequencing data from multiple experimental runs can be used together to detect CNVs as it eliminates the batch biases. CoNIFER can robustly detect rare CNVs and estimate the copy number of duplicated genes up to approximately 8 copies with current exome capture kits [167].

The other widely used tool for CNV detection from exome sequencing data is ExomeDepth [161]. It uses a robust beta-binomial distribution for the read count data. Unlike CoNIFER, ExomeDepth build an optimized reference set using the beta-binomial model in order to maximize the power to detect CNVs [161]. Briefly, for each test sample, ExomeDepth ranks the remaining samples by order of correlation with the test sample. Next, the reference set is generated by adding the samples sequentially [161]. This is the main difference between the approach used in CoNFIER and ExomeDepth. CoNFIER tries to eliminate biases in exome data whereas ExomeDepth creates a reference sample and performs a comparison between reference sample and the test sample. The power of ExomeDepth highly depends on how good the samples are correlated. The high correlation among the samples can be generated when samples are prepared in almost exactly the same way. In this study, a novel method has been introduced, which was compared to ExomeDepth and CoNFIER (see Chapter 4).

Hidden Markov models (HMMs) are useful natural framework in CNV detection that can segment genomic data with a discrete number of states [171]. HMMs are a statistical model that can be used to determine an unknown sequence of states based upon a sequence of observations. In other words, in HMM, the sequence of states is hidden and can only be inferred through a sequence of observed random variables. HMMs have Markov property, which means that each new state of a sequence is only dependent upon the previous state. A change from any one state to another is described by a matrix of transition probabilities. In addition to a state transition probability distribution, each state in a HMM has an emission probability distribution modeling the observed variable as a function of a particular hidden state.

Firstly, HMMs optimize the model parameters meaning the emission and transition probability distributions, to best describe the observed sequence of variables. Next, using a dynamic programming approach, HMMs can infer the most probable sequence of hidden states. An example of HMM is shown in Figure 2.8, constructed by a random process $\{X_m, Y_n\}$. Here, X_m are the hidden states and the transition probability is given by $P(X_{m+1}|X_m)$, which determines the probability of the state of $m+1$ based on the state of m. Another important component of a HMM is sequence of observations, meaning Y_n. Each hidden state generates an observation with specific probability, $P\{Y_n|X_m\}$, which is called the emission probability. The use of HMMs in this study is described in Chapter 4.

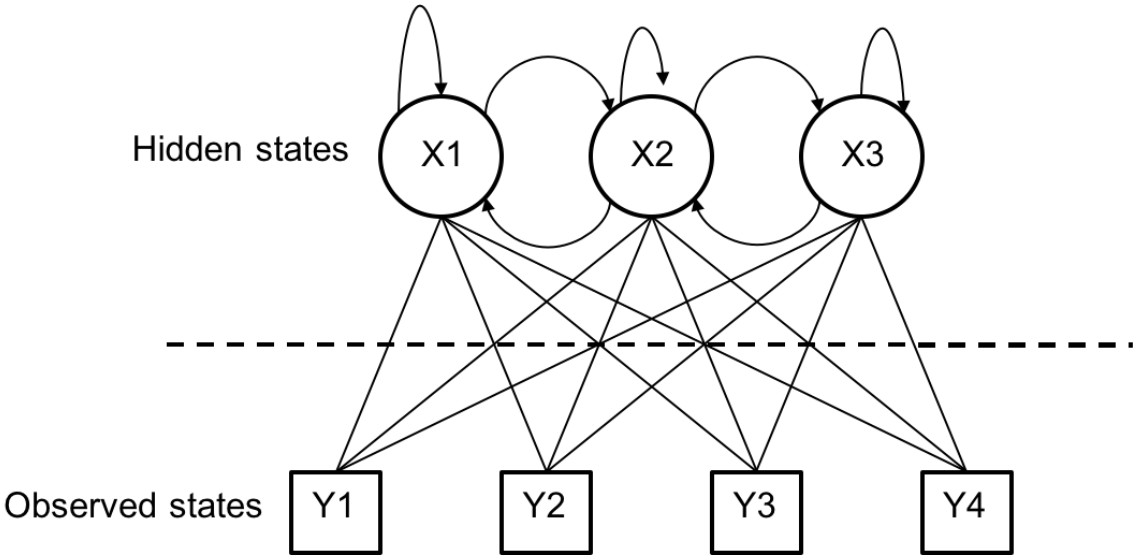


Figure 2.8: Example of a Hidden Markov Model specification. Here, X is a hidden state and Y is a observed state. The transition probabilities are shown by curved arrows and emission probabilities are shown by lines.

2.2.3 Analysis of ChIP-seq Data

Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing allows genome-wide identification of protein-DNA interactions such as transcription factor bindings, transcriptional co-factors binding, RNA polymerases binding and chemical modifications of histone proteins [118–122]. In general, the first step in the ChIP-seq analysis is read mapping (Chapter 2.2.1), followed by the peak calling step, which aims to identify the genome-wide binding sites of a protein. In this study, ChIP-seq data of histone marks and transcription factor has been used. For histone

marks, in addition to the peak calling step, I performed an analysis to check the enrichment pattern of histone marks around the transcription start sites (TSS). All the steps used for ChIP-seq data analysis has been described in the following sections.

2.2.3.1 Identification of Genome-wide Binding Events

After mapping, the most common step is the peak calling to identify the genome-wide binding sites of a protein. In the past, various tools have been developed to find peaks from the ChIP-seq data [172–176]. In general, binding sites or peaks are the regions of a genome where sequence reads are significantly enriched as compared to the control. Before defining a region as a peak, distinct steps are carried out such as read shifting and background estimation. During the mapping step, ChIP-seq reads can align to either the sense or antisense strand and therefore, location of mapped reads form two peaks. Next, the reads are shifted towards the centre to determine the most likely location involved in protein binding (Figure 2.9). The shift parameter is determined by the fragment size generated in the ChIP-seq library preparation. Interestingly, Model-based analysis of ChIP-seq (MACS) can empirically model the shift size of ChIP-seq reads without any prior knowledge [177].

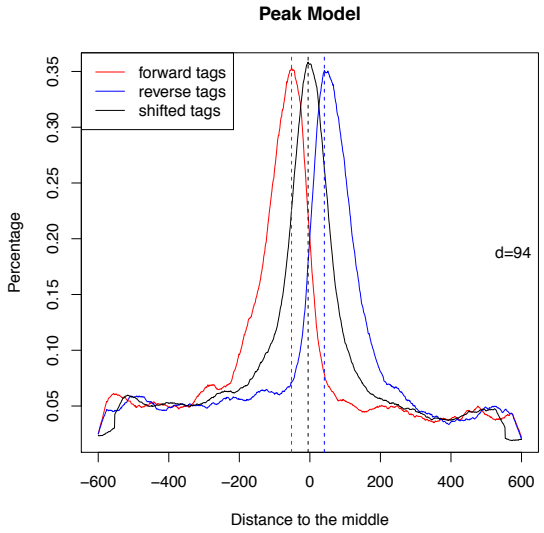


Figure 2.9: Shifting size model is generated by MACS using MyoD ChIP-seq data from undifferentiated C2C12 cells. Here d is the estimated fragment size and reads (tags) are shifted by $d/2$.

A peak in treated sample is identified by comparing to the control sample (input control or IgG control). Both the samples are processed in the same way for the

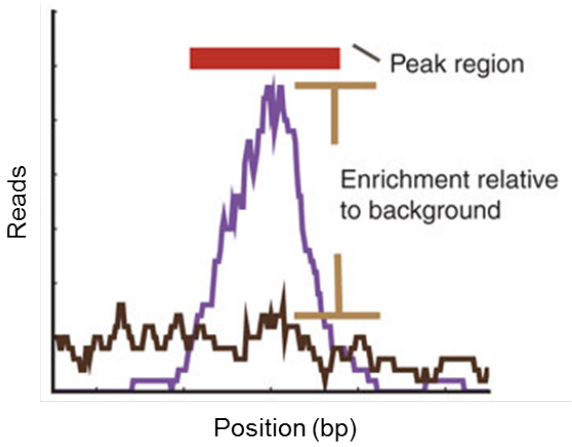


Figure 2.10: Example of an enriched region as compare to the background. The purple line shows the read enrichment in the treatment sample and the brown line shows the read enrichment in the control sample. Figure taken from Pepke *et al.* [178] and modified.

comparison. The reference genome is divided into the windows and read enrichment is compared between the treatment and control sample (Figure 2.10). More often, a p-value is calculated to identify a peak using a statistical model. Other strategy is to use the read enrichment and fold change over the background, but these do not provide statistical significance values. Thus, the Poisson distribution has frequently been used to derive significantly enriched windows. The most widely used tool MACS uses a dynamic parameter, λ_{local}, to compensate the local fluctuations and is defined as

$$\lambda_{local} = max(\lambda_{BG},\ \lambda_{1k},\ \lambda_{5k},\ \lambda_{10k})$$

where λ_{BG} is a uniform estimation for the whole genome, $\lambda_{1k},\ \lambda_{5k},\ \lambda_{10k}$ are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample [175]. MACS uses Poisson distribution which defines the probability of finding a number of k reads mapped to the window as

$$Pr(X = k) = \frac{\lambda_{local}^{k}\ e^{\lambda_{local}}}{k!}$$

In general, peaks are classified into the point source, broad source and mixed source depending on the protein immunoprecipitated [179] (Figure 2.11). As the name suggests, point sources are the narrow peaks generated from sequence-specific transcription factors whereas broad peaks are generated from chromatin marks and cover larger regions [179]. Mixed source peaks are the enriched regions of a range of

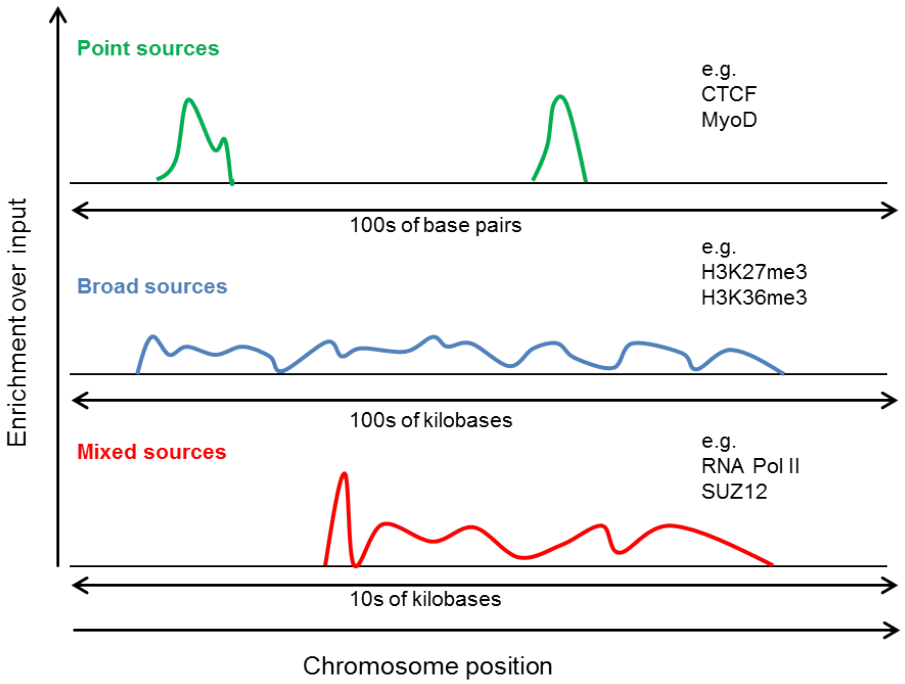


Figure 2.11: Different types of peaks in ChIP-seq experiment. Figure drawn on the basis of Sims *et al.* [179].

sizes which are generated by proteins such as RNA polymerase II and transcriptional repressor CTCF [179]. It remains difficult to call peaks for broad source and mixed source factors since the length of the enriched region is several kilobase (kb). Nevertheless, new methods are emerging to identify such regions. For example, MACS2 (an updated version of MACS) is specifically designed to process mixed signal types [127, 175, 180].

2.2.3.2 Annotation of ChIP-seq Peaks

One of the important step after peak calling is to summarize the location of the peaks in the genome. This step can be used as a validation criterion for certain chromatin-associated modifications and proteins [127]. For example, using prior knowledge, we can confirm if a particular sequence-specific transcription factor preferentially binds near transcription start site (TSS). In this study, I used several definitions to annotate the peaks (Figure 2.12). If a peak overlaps two regions, it was annotated for both of them. More often, researchers are interested in associating peaks with the genes, therefore in this study peaks were assigned to the genes according to the criterion suggested by Schlesinger et al [37]. Peaks were assigned to the genes if they are located within 10 kb upstream of the TSS or in the transcribed region.

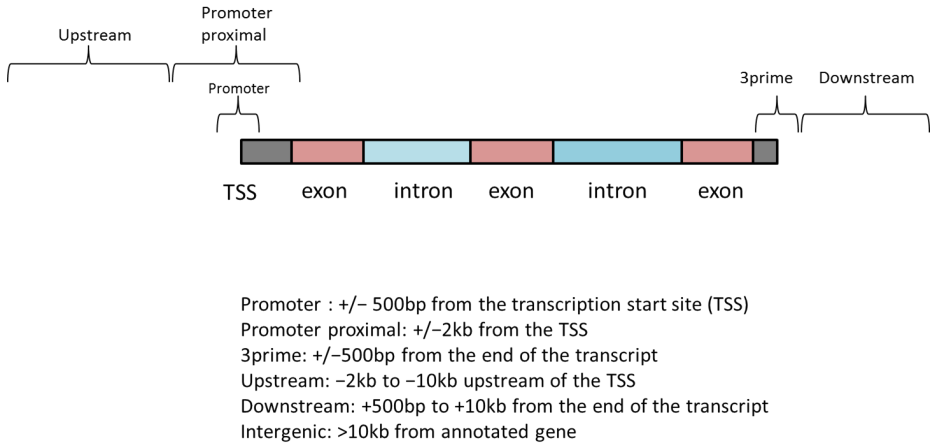


Figure 2.12: Defined regions for annotating ChIP-seq peaks.

2.2.3.3 Discovery of Sequence Binding Motifs

Sequence-specific transcription factors preferentially binds to a short DNA sequence, which is expected to be enriched in ChIP-seq peaks. Therefore, when a motif of the protein is already known, this step can be used as a proof of principle of a successful experiment. Moreover, if the motif is not known, discovery of a centrally located motif can lead to the identification of DNA-binding motifs of other proteins that bind in complex, which highlights the mechanism of transcriptional regulation [181]. A commercially available database "TRANSFAC" provides the experimentally-proven binding sites of eukaryotic transcription factors [182]. These motifs are stored as position weight matrices (PWMs), also called as position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM), and is the most common way to represent the motifs. In the past, various tools have been introduced to identify over-represented motifs in the ChIP-seq peaks [183–185]. For example, MATCH tool uses the matrix library collected in TRANSFAC database and calculates two score values: the matrix similarity score (MSS) and the core similarity score (CSS). These two scores measure the quality of a match between the sequence and the matrix, which ranges from 0.0 to 1.0, where 1.0 denotes an exact match [182, 185]. The core of each matrix is defined as the first five most conserved consecutive positions of a matrix and CSS is calculated for all pentanucleotides and prolonged at both ends, so that it fits the matrix length [182, 185]. Both scores, MSS and CSS, are calculated in the following way

$$MSS\ or\ CSS = \frac{Current - Min}{Max - Min}$$

where, *Current* is the frequency of nucleotide B to occur at the position i of the matrix ($B \in \{A, T, G, C\}$), *Min* is the frequency of the nucleotide which is rarest in position i in the matrix and *Max* is the highest frequency in position i. Next, over-represented transcription factor binding sites that are statistically significant are identified assuming a binomial distribution. In the output, these tools give a sequence logo, which is a common graphical representation for a matrix [186]. An example of a frequency matrix and its sequence logo is shown in Figure 2.13.

	1	2	3	4	5	6	7
A	15	4	41	36	7	19	3
C	11	35	1	2	29	14	22
G	10	2	1	4	6	7	15
T	7	2	0	1	1	3	3

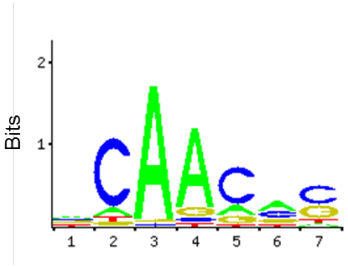


Figure 2.13: Example of a frequency matrix (left) and its sequence logo (right). Data randomly taken from open-source JASPAR database [187].

Discovery of binding motifs using MATCH, will return only the known motifs collected in TRANSFAC database and therefore in this study, I also used Regulatory Sequence Analysis Tools (RSAT) for the detection of novel motifs, known as *de novo* motif discovery [188, 189]. Using RSAT, oligo-analysis was performed for detecting over-represented oligonucleotides. Additionally, RSAT can also compare the discovered motifs with databases like JASPAR [187]. The advantage of using two tools is that we can focus on high confident motifs, which are discovered by both of them.

2.2.3.4 Gene Ontology Enrichment Analysis

One of the most common analysis of a gene set is Gene Ontology (GO) Enrichment Analysis. It provides core biological knowledge representation of a gene list. The Gene Ontology project provides an ontology which covers three domains - "Molecular Function", "Biological Process", and "Cellular Component" [190, 191]. GO analysis was conducted using the DAVID functional annotation tool [192]. The tool calculates an EASE Score using a modified Fisher's exact p-value. Here is an example of how functional annotation tool of DAVID calculates a p-value. Consider a human genome as background with 30,000 genes. Out of these, 40 genes are involved in vasculature development. On the other hand, a gene list provided by the user has 3 out of 300 genes involved in vasculature development. Using a Fisher's exact test, we can test the hypothesis if 3/300 is more than random chance comparing to the human

	User genes	Background
In pathway	3	40
Not in pathway	297	29,960

Table 2.1: A 2x2 contingency table built for Fisher's exact test.

background of 40/30000 (Table 2.1). Using 3 for the test, we get a p-value of 0.008, which is significant (p-value $\leq$ 0.01) suggesting that user's gene list is enriched for vasculature development. Functional annotation tool of DAVID uses 3-1 (instead of 3) to make the test more stringent and conservative. Therefore, p-value or EASE Score is 0.06 which is not significant. However, it is worthy to note that annotations in the DAVID database are not regularly updated. Therefore, I used Mouse Genome Informatics (MGI) database to validate the results of DAVID database [193]. Specific GO terms were downloaded from MGI database and using two-sided Fisher's exact test, I confirmed if these terms are indeed significantly enriched or not.

2.2.3.5 Enrichment Profile Around the Transcription Start Sites

In the past, several studies have shown the importance of plotting the ChIP-seq signal or enrichment around transcription start site (TSS), specially for histone marks [194–196]. Therefore, in addition to peak calling approach, I used the enrichment profile of histone marks around the TSS. ChIP-seq signal can be calculated around the TSS, for example, 2 kb upstream and 4 kb downstream of TSS. The resulting region of 6 kb length can be divided in 100 bp long non-overlapping windows. For each window, the total signal is calculated on the basis of mapped reads (see Chapter 5). Moreover, based on the ChIP-seq signal, previous studies have suggested to cluster the genes into subgroups using k-means clustering [194, 195]. Genes in the same cluster have the similar pattern or profile of histone signal around the TSS and each cluster has a unique enrichment profile, which is different from the rest. K-means is a non-hierarchical clustering method that is used to classify data into groups of genes without specifying relationships between genes in a cluster. This method requires predetermined number of clusters, which can be defined based on the visual inspection. The first step is the initialization, in which an average (centroid) is calculated for each cluster and then, genes are reassigned to the different clusters depending on which centroid it is closer. Calculation of centroid and re-grouping is

performed in an iterative manner. In this study, k-means clustering was performed in R (v3.0.2) using the "kmeans" function and 100,000 iterations.

2.2.4 Analysis of RNA-seq Data

RNA-seq is commonly used to deduce and quantify the gene expression. Moreover, different samples can be compared to each other to find differentially expressed genes or isoforms. Apart from this, RNA-seq is used to identify differential alternative splicing events. In this study, all three applications have been used, meaning the quantification of gene expression, the identification of differentially expressed genes and the detection of differential exon usage. Analysis of RNA-seq data for above mentioned applications need sophisticated tools with advanced statistical models to reduce the false positives. Some of these tools have been discussed in the following sections.

2.2.4.1 Quantification of Gene Expression

After mapping RNA-seq reads to a reference sequence, expressed genes can be identified using the mapped reads. In general, more reads will be mapped to long genes and less reads will be mapped to short genes. Therefore, to quantify and compare gene expression within the sample needs a normalisation step according to the gene length. Moreover, before comparing genes in different samples, data needs to be normalise for differences in library size or sequencing depth [197]. The most common normalisation method is to calculate RPKM (reads per kilobase per million mapped reads) for single-end reads or FPKM (fragments per kilobase per million mapped fragments) for paired-end reads. This approach facilitates the comparison between genes within a sample and between the samples as it normalises the data for library size and also for the gene length [114, 197]. RPKM of a particular gene is defined as

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

where C is the number of mappable reads that fell onto the gene exons, N is the total number of mappable reads in the experiment and L is the sum of the exon length in base pairs [114, 197]. In contrast to quantifying gene expression, estimation of isoform expression remains difficult because more often reads are mapped to multiple isoforms. To solve this issue, Trapnell *et al.* developed an algorithm, Cufflinks, that can estimate the abundance of isoforms or transcripts by probabilistically assigning reads to the isoforms [198]. The probability that a fragment originates from transcript

t and the probability of selecting a fragment from transcript t are denoted by β_g and γ_t respectively. These parameters are estimated from a likelihood function and the abundance of a transcript $t \in$ gene g is given in FPKM units

$$Cufflinks\ FPKM = \frac{10^9\ \beta_g\ \gamma_t}{l(t)},$$

where *l(t)* is an adjusted length of transcript t [198]. Cufflinks is a part of Tuxedo tools and is widely used for assembling and quantifying the gene and/or isoform expression [199].

Genes or transcripts, even with only one mapped read will produce a signal and the FPKM or RPKM value will be greater than zero. The major challenge after calculating FPKM values is to determine whether the transcript is functional in a cell/tissue or not. Hebenstreit *et al.* showed (Figure 2.14) that the distribution of RPKM (or FPKM) values divides the expression level in two groups of genes, meaning lowly expressed (LE) and highly expressed (HE) genes [200]. The study suggested that the RPKM value of at least one seems a fair cutoff to define putatively functional genes and moreover, they found that one RPKM corresponds to an average of roughly one transcript per cell [200]. Therefore, in the present study, I have used this cutoff to define putatively functional genes in our data.

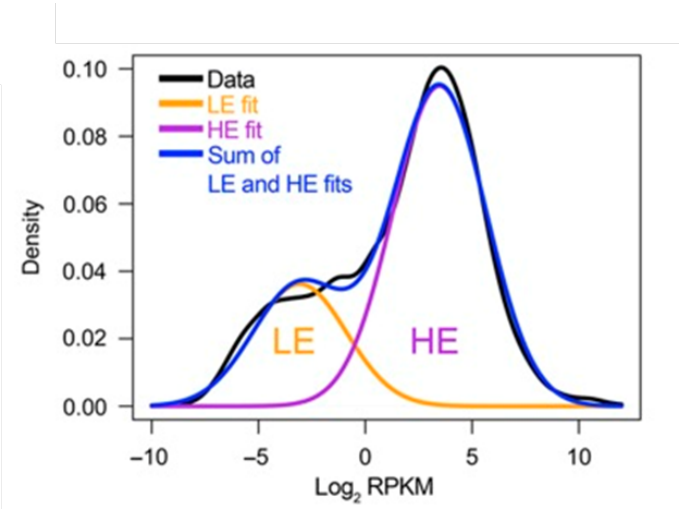


Figure 2.14: Expectation-maximization-based curve fitting of RNA-seq data shows the two groups of genes i.e. lowly expressed (LE) and highly expressed (HE) genes. Figure taken from Hebenstreit *et al.* [200].

2.2.4.2 Identification of Differentially Expressed Genes

After transcriptome profiling, one major interesting question is the discovery of differentially expressed genes across different conditions (e.g., patients versus healthy individuals, right ventricle tissue versus left ventricle tissue, undifferentiated cells versus differentiated cells). In general, to define a gene as differentially expressed, the variation between the groups (e.g., patients versus healthy individuals) should be significantly large as compared to the within each group (e.g. within patient cohort). To estimate the better variation, we need biological replicates for each condition but it remains difficult to define a standard number of replicates. In general, more replicates will lead to better estimation of variation but it increases the overall cost of the experiment. One simple way for the discovery of differentially expressed genes without replicates is to use the fold change between two conditions. For example, if a gene A in condition X has a FPKM value of 10 and in condition Y has a FPKM value of 5, the fold change will be 2 and depending on the cutoff, we can define if gene A is differentially expressed or not. On the other hand, it is always better to use a statistical model for replicates to identify differentially expressed genes. The statistical computation becomes more complex at transcript resolution. To address this issue, Trapnell *et al.* introduced a tool, Cuffdiff 2, by modeling variability in the number of fragments generated by each transcript across replicates [201].

It remains difficult to choose a model that controls for variability in technical and biological noise [202]. One natural choice and commonly used model for fragment count is the Poisson model, in which the variability is estimated by calculating the mean count across replicates but it does not provide enough flexibility, meaning more variability exists than can be explained by the model (overdispersion) and does not address the issue of count uncertainty (reads map ambiguously to different transcripts) [203–205]. Previously, it has been observed that overdispersion in RNA-seq experiments increases with expression and therefore, the negative binomial distribution has been proposed as a means of controlling for it but the issue of count uncertainty is not addressed [203, 206]. Cuffdiff 2 algorithm captures uncertainty in a transcript's fragment count as a beta distribution and the overdispersion in this count with a negative binomial, and mixes the distributions together (i.e. beta negative binomial distribution) [201].

In contrast to FPKM based strategy, Love *et al.* presented a method, DESeq2, for differential analysis of count data [207]. They use shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates [207]. Read counts in a matrix K_{ij} are modeled using a negative binomial distribution with mean

μ_{ij} and dispersion σ_i. DESeq2 deals with the experiments with small number of replicates (two or three) that leads to highly variable dispersion estimates for each gene and therefore, assume that genes of similar average expression level have similar dispersion [207]. The shrinkage of maximum a posteriori dispersion estimates toward the fitted values is shown in Figure 2.15.

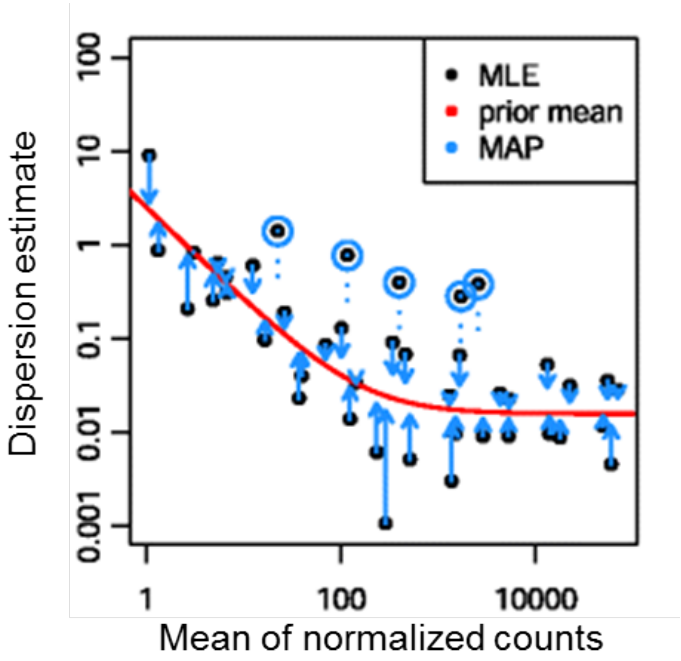


Figure 2.15: Shrinkage estimation of dispersion using DESeq2. Figure taken from Love *et al.* [207]. MAP, maximum a posteriori; MLE, maximum-likelihood estimate.

To make the explanation easier to understand, the aforementioned strategies have been described with an example of a single gene. But more often, we test thousands of gene for the differential expression. In other words, thousands of p-values are calculated using the same test for different genes. This leads to the multiplicity problem, in which thousands of hypothesis are tested simultaneously and the chance of false positives significantly increases. For example, performing the same test 10,000 times, one would expect 10,000 * 0.01 = 100 of them to have a p-value $\leq$ 0.01, even in a completely random situation. Therefore, in order to reduce the false positives, an adjustment to the p-values is needed. Over the past several years different methods have been introduced to address this issue. For example, Bonferroni adjustment is commonly used but this method is too conservative [208, 209]. Therefore, Cuffdiff 2 and DESeq2 implements a FDR (false discovery rate) adjusted p-value method using Benjamini and Hochberg procedure [210].

2.2.4.3 Detection of Differential Exon Usage

It is known that more than 90% of multiexon genes in human, undergo alternative splicing [211, 212]. The inclusion or exclusion of different exons in a mature RNA leads to the translation of different proteins. Therefore, one gene could give rise to multiple proteins and potentially, with different function. Altered RNA splicing is implicated in many human diseases [213–215], therefore it is important to study this process. RNA-seq provides the opportunity to study alternative splicing but requires sophisticated computational methods to analyze the data. More often, we are interested in comparing different conditions (e.g. knockout mice versus wildtype mice) and ask, if a exon is differentially used between the conditions. As in the analysis of differentially expressed genes, it is difficult to choose a model that controls for variability in biological noise for detection of differential exon usage. Wang *et al.* used the 2 x 2 contingency tables of read counts and applied Fisher's exact test to identify differentially used exons [211]. Recently, this method was extended by considering the read coverage for the alternative exon and its corresponding gene [216]. Firstly, a p-value is calculated between the two conditions using junction read counts by Fisher's exact test. Next, using the read coverage for the alternative exon and its corresponding gene, a second p-value is calculated by Fisher's exact test. The two p-values are combined to get an adjusted p-value using a weighted arithmetic equation [216]

$$Pvalue_{adjusted} = w * Pvalue_{first} + (1 - w) * Pvalue_{second},$$

where w is the weight of the $Pvalue_{first}$, whose value depends on the size of an alternative exon and read length [216]. This method has been implemented in Java and called as Alternative splicing detector (ASD). The advantage of ASD is that it can compare the two conditions without using replicates. To address the issue of biological variation, Anders *et al.* presented a method, DEXSeq, which models the read count using negative binomial (NB) distribution [217]

$$K_{ijl} \sim NB\ (mean = s_j \mu_{ijl},\ dispersion = \alpha_{il}),$$

where K_{ijl} is the number of reads overlapping counting bin l of gene i in sample j. μ_{ijl} is the predicted mean, s_j is the size factor, which accounts for the sequencing depth of sample j, and α_{il} is a measure of the distribution's spread [217]. The advantage of DEXSeq is that it includes the information from biological variation using replicates.

Over the past few years, it has been suggested that estimation of percent-spliced-in (PSI, Ψ) captures more accurately the local information related to splicing of each particular exon [218–220]. Moreover, Guo *et al.* showed the use of difference in PSI value of exons between two conditions (i.e. $\triangle$PSI), without using replicates [221]. PSI metric takes the advantage of junction reads, which supports the inclusion or exclusion of an exon under consideration. Ψ is defined as

$$\Psi = \frac{a+b}{a+b+2c},$$

where $a+b$ is the number of reads supporting the inclusion of an exon and c is the number of reads supporting the exclusion of an exon. The factor of two in the denominator accounts for the fact that there are twice as many mappable positions for reads supporting exon inclusion as exon exclusion [221]. In general, PSI value of 1 means that the exon is fully included (100%). Figure 2.16 shows the junction reads, which support the inclusion (a and b) and exclusion (c) of an exon.

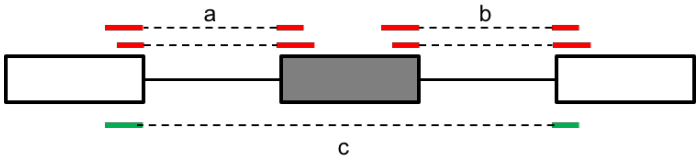


Figure 2.16: Illustration of percent-spliced-in (PSI) metric. The exon filled in grey is assumed to be under consideration. $a+b$ is the number of reads supporting the inclusion of the exon and c is the number of reads supporting the exclusion of the exon.

During the comparison between the two conditions, PSI values are calculated for each exon in both the conditions. Therefore, each exon will have two PSI values and difference between them can be calculated. In this study, based on PSI values, a pipeline has been generated to detect differential exon usage by comparing the two conditions (see Chapter 6).

Chapter 3

Project-Related Datasets

Using the aforementioned methods (i.e. targeted resequencing, ChIP-seq and RNA-seq), multiple experimental datasets were generated by the Sperling group (Cardiovascular Genetics at the ECRC, Charité Campus Berlin-Buch) to investigate various components during heart and skeletal muscle development and disease. All the described datasets are computationally analyzed within this thesis. Unless stated otherwise, high-throughput sequencing was performed by Bernd Timmermann's (Next Generation Sequencing Service) at the Max Planck Institute for Molecular Genetics, Wei Chen' group (Scientific Genomics Platform) at the Berlin Institute for Medical Systems Biology, and ATLAS Biolabs GmbH.

3.1 DNA-seq Data From Patients with Tetralogy of Fallot

Targeted resequencing was performed for eight patients with Tetralogy of Fallot (TOF), all of which are unrelated sporadic cases with a well-defined coherent phenotype and no further anomalies. Genomic DNA (gDNA) was extracted from whole blood of five TOF patients and from right ventricle of three TOF patients (Table 3.1). These patient samples were collected in collaboration with the German Heart Center Berlin and the National Registry of Congenital Heart Disease in Berlin. The quality of gDNA was assessed on agarose gel and spectrophotometer. Three to five µg of gDNA were used for Roche NimbleGen sequence capturing using 365K arrays. For array design, 867 genes and 167 microRNAs (12,910 exonic targets representing 4,616,651 target bases) were selected based on knowledge gained in various projects [37, 99, 222, 223]. DNA enriched after NimbleGen sequence capturing was sequenced

using the Illumina Genome Analyzer (GA) IIx (36 bp paired-end reads). Sequencing was performed by ATLAS Biolabs (Berlin) according to the manufacturer's protocols.

Sample	Gender	Age category	Source for lib prep	Number of reads	Number of read pairs	Mean phred quality score	Median coverage	Mean coverage	Target bases with ≥10x coverage
TOF-01	Male	1-3 years	RV	31,942,782	15,971,391	33.3	40	47	93.85%
TOF-02	Female	1-3 years	RV	26,970,680	13,485,340	32.7	66	76	97.70%
TOF-18	Female	Infant	RV	25,476,308	12,738,154	35.4	71	80	98.35%
TOF-23	Male	Infant	Blood	20,885,192	10,442,596	35.0	60	69	97.41%
TOF-24	Male	Infant	Blood	25,483,166	12,741,583	34.7	51	58	96.72%
TOF-25	Male	Infant	Blood	30,551,674	15,275,837	34.6	84	92	98.91%
TOF-26	Female	Infant	Blood	27,878,750	13,939,375	34.7	75	84	98.34%
TOF-27	Female	Infant	Blood	24,118,022	12,059,011	34.6	78	90	98.00%

Table 3.1: Sample information, number and quality of 36 bp paired-end reads obtained from targeted resequencing in TOF patients using Illumina's Genome Analyzer IIx platform.

3.2 DNA-seq Data From HapMap Samples

Exome sequencing data from eight HapMap individuals (NA18507, NA18555, NA18956, NA19240, NA12878, NA15510, NA18517, NA19129) were analyzed [224]. The exomes were captured using Roche NimbleGen EZ Exome SeqCap Version 1 and sequencing was performed using an Illumina HiSeq 2000 platform with 50 bp paired-end reads (Table 3.2). The exome sequence data was downloaded from the Sequence Read Archive (SRA) at the NCBI (SRA039053).

Sample	Gender	Super Population	Number of read pairs	Number of reads
NA18507	Male	African	90,320,349	180,640,698
NA18555	Female	East Asian	109,683,131	219,366,262
NA18956	Female	East Asian	78,615,001	157,230,002
NA19240	Female	African	79,085,299	158,170,598
NA12878	Female	European	85,493,722	170,987,444
NA15510	Unknown	Unkown	80,579,037	161,158,074
NA18517	Female	African	89,101,813	178,203,626
NA19129	Female	African	83,554,193	167,108,386

Table 3.2: Sample information and number of 50 bp paired-end reads from HapMap project.

3.3 ChIP-seq Data of Transcription Factor MyoD and Histone Modifications From C2C12 Skeletal Muscle Cells

C2C12 cells are murine myoblasts, which were originally derived from adult dystrophic mouse muscle [225] and are a useful model to study myogenic differentiation. We performed ChIP-seq for H3K4me2, H3K4me3 and MyoD in undifferentiated (Undiff) and differentiated (Diff) C2C12 cells. To induce differentiation, cells were cultured with Dulbecco's modified Eagle's medium and 2% horse serum (Biochrom) and maintained for 48 hours, when more than 90% of the cells had fused into myotubes (Diff C2C12 cells). ChIP was performed with the MAGnify Chromatin Immunoprecipitation System (Life Technologies, 49-2024) following the manufacturers instructions with modifications. Sonication was performed using the Biorupter UCD300 (Diagenode) to obtain chromatin fragments of approximately 100-300 bp. The following antibodies were used for ChIP: anti-H3K4me2 (Abcam ab7766), anti-H3K4me3 (Abcam ab8580), and anti-MyoD (Santa Cruz, sc-760). Sequencing libraries were prepared using the NEXTflex ChIP-Seq Kit (Bio Scientific, 5143) according to an in-house (Sperling lab) modified protocol. The libraries were 51 bp single-end sequenced on an Illumina HiSeq 2000 platform. Base calling was performed with the Illumina Casava pipeline version 1.8.0. Initial sequencing quality assessment was based on data passing the Illumina Chastity filter. Sequencing of DNA libraries resulted in approximately 29-66 million reads per sample (Table 3.3).

Samples	Lane	Stage	Number of reads	Number of reads mapped uniquely
H3K4me2	4	Undiff	23,731,636	19,244,892
	5	Undiff	23,486,840	19,040,146
H3K4me3	4	Undiff	31,793,027	22,480,368
	5	Undiff	31,599,195	22,332,465
MyoD	3	Undiff	30,581,335	21,370,215
Input	6	Undiff	33,423,759	24,036,549
	7	Undiff	32,484,546	23,481,460
H3K4me2	4	Diff	33,150,496	26,894,013
	5	Diff	32,997,808	26,758,161
H3K4me3	4	Diff	27,906,380	19,207,470
	5	Diff	27,751,745	19,093,615
MyoD	3	Diff	28,634,899	19,834,717
Input	6	Diff	26,982,855	19,148,808
	7	Diff	27,554,807	19,645,534

Table 3.3: Overview of total number of reads generated per sample in ChIP-seq during myogenic differentiation. Sequence reads are single-end 51 bp long and some DNA libraries were sequenced on multiple lanes.

3.4 RNA-seq Data From C2C12 Skeletal Muscle Cells

RNA-seq data were used from the ENCODE project [127] for undifferentiated C2C12 cells and differentiated C2C12 cells (60 h timepoint). To induce differentiation, cells were cultured with Dulbecco's modified Eagle's medium and 2% donor equine serum (HyClone) and maintained for 60 hours. The libraries were 75 bp paired-end sequenced on an Illumina Genome Analyzer II platform (204 and 185 million paired-end reads in undifferentiated and differentiated C2C12 cells, respectively). The data were downloaded from the Sequence Read Archive (SRA) at NCBI with accession numbers SRR496442 (undifferentiated C2C12 cells) and SRR496443 (differentiated C2C12 cells).

3.5 RNA-seq Data From Dpf3 Knockout Mice

In previous studies, the Sperling laboratory identified the chromatin remodeling factor Dpf3, the expression of which was significantly up-regulated in the right ventricle of TOF patients [99, 100]. Moreover, in-house (Sperling lab) experiments showed that Dpf3 interacts with splicing factors, which suggests its potential role in splicing. Therefore, to study the role of Dpf3 in splicing, Dpf3 knockout mice were generated and mRNA sequencing (mRNA-seq) was performed. Using knockout (KO) and wildtype (WT) mice, we extracted mRNA from three tissues; namely, right ventricle (RV), left ventricle (LV) and skeletal muscle (SM). The strand-specific libraries were prepared using the "ScriptSeq RNA-seq library preparation kit" from Illumina and paired-end sequencing was performed on an Illumina HiSeq 2000 platform with a read length of 50 bp. The libraries were sequenced on multiple lanes (Table 3.4).

Dpf3 KO mice
(2 males and 2 females)
12 weeks old

Sample	Lane	Number of reads	Number of read pairs
RV	2	42,631,450	21,315,725
	3	42,277,536	21,138,768
	4	42,409,714	21,204,857
LV	2	42,029,952	21,014,976
	3	41,770,368	20,885,184
	4	41,951,434	20,975,717
SM	2	44,893,392	22,446,696
	3	44,508,418	22,254,209
	4	44,751,428	22,375,714

Dpf3 WT mice
(2 males and 2 females)
12 weeks old

Sample	Lane	Number of reads	Number of read pairs
RV	2	46,630,042	23,315,021
	3	46,257,634	23,128,817
	4	46,431,508	23,215,754
LV	2	53,350,132	26,675,066
	3	52,956,668	26,478,334
	4	53,242,586	26,621,293
SM	2	48,132,018	24,066,009
	3	47,848,484	23,924,242
	4	48,071,502	24,035,751

Table 3.4: Overview of total number of reads generated per sample in Dpf3 knockout (KO) and wildtype (WT) mice. Sequence reads are paired-end 50 bp long and strand-specific libraries were sequenced on multiple lanes. RV, right ventricle; LV, left ventricle; SM, skeletal muscle.

Chapter 4

Outlier-Based Identification of Copy Number Variations Using Targeted Resequencing in a Small Cohort of Patients with Tetralogy of Fallot

4.1 General Purpose

Copy number variations (CNVs) are one of the main sources of variability in the human genome. Many CNVs are associated with various diseases including cardiovascular disease. These copy number changes are usually defined to be longer than 500 bases, including large variations with more than 50 kilobases [226, 227]. Previous studies have identified CNVs in large cohorts of non-syndromic patients with Tetralogy of Fallot (TOF). All three studies used SNP arrays to identify CNVs [55–57]. Observing the overlap between these studies with hundreds of cases revealed only one locus (1q21.1) affected in 11 patients (Figure 4.1), which underlines the heterogeneous genetic background of non-syndromic TOF. In this study, we aimed to identify copy number alterations in a small cohort of non-syndromic TOF patients based on targeted resequencing data. We developed a novel CNV calling method to identify individual/personalised disease-relevant CNVs. The method is based on outlier detection using Dixon's Q test and assessment of outliers using a Hidden Markov Model (HMM). For evaluation, we applied our method to a small cohort of HapMap samples and compared it to results obtained by ExomeDepth and CoNIFER. Subsequently, our method and CoNIFER were used to detect CNVs in the TOF patients. For this project, we analyzed targeted resequencing data from 8 patients with TOF and exome

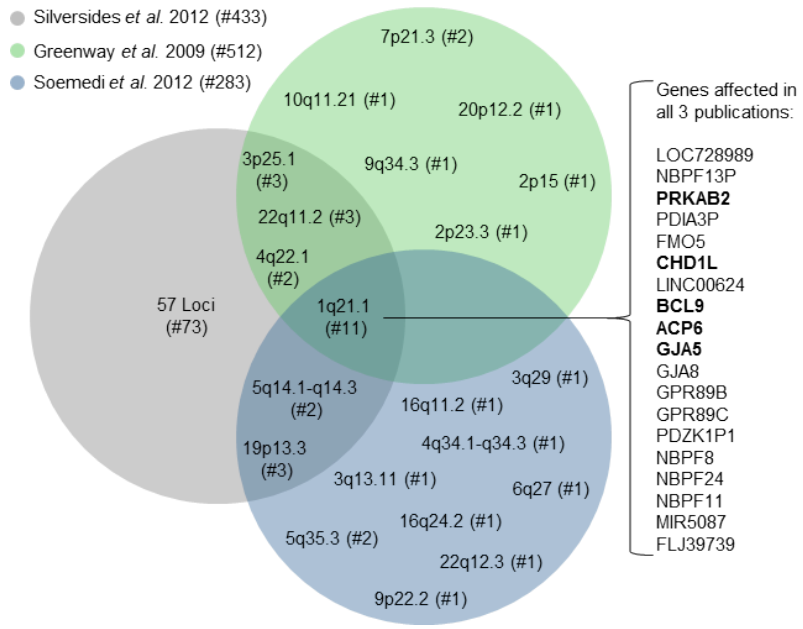


Figure 4.1: All three studies are based on SNP arrays. Loci with detected CNVs are depicted according to their respective cytoband. For 1q21.1, which was identified in all three studies, the RefSeq genes that are affected in at least one patient in each of the publications are listed in the order of their genomic position.

sequencing data from 8 HapMap individuals (Chapter 3.1 and 3.2).

4.2 Novel Outlier-Based Copy Number Variation Calling Method

Our CNV calling method was developed for exome or targeted resequencing data of small sets of samples (at least 3 and at most 30) assuming that the bias in the captured regions is similar in all samples enriched and sequenced with the same technology. Based on a heterogeneous genetic background in the cohort, it was further assumed that a unique disease-related copy number change is only present in very few samples.

4.2.1 Calculation of Copy Number

In the first step, we calculated the copy number values for each sample separately. The sequenced reads were mapped to the targeted regions of the reference genome using BWA (v0.5.9) in paired-end mode ('sampe') with default parameters. During targeted resequencing, often up- and downstream of the targeted regions (usually exons) are also captured. Therefore, the regions were extended by 35 bp (read length minus one

base pair) to correctly capture the coverage at the start and end of a region. After mapping, the extended regions with their mapped reads were joined chromosome-wise and the tool mrCaNaVaR (v0.34) was used to split the joined regions into non-overlapping windows of 100 bp in length. The copy number value C for each window $W \in \{1, ..., n\}$ of a sample $S \in \{1, ..., n\}$ was then calculated by mrCaNaVaR using the following formula

$$C_W^S = \frac{Number\ of\ reads\ mapped\ to\ W}{Average\ number\ of\ reads\ mapped\ over\ all\ windows} * 2.$$

4.2.2 Identification of Outliers Using Dixon's Q Test

Dixon's Q Test was introduced in 1950 for the analysis of extreme values and for the rejection of outlying values [228]. We used the formulas for r_{10} and r_{20}, also known as type10 and type20 in the R package 'outliers' (v0.14) [229]. For this test, firstly we have to arrange the values in ascending order $x_1 < x_2 < ... < x_n$. Then, the experimental Q-value (Q_{exp}) is calculated. The equations for calculating Q-values for r_{10} and r_{20} are given below

$$r_{10} = \frac{x_2 - x_1}{x_n - x_1}, for\ a\ single\ outlier\ x_1,$$

$$r_{20} = \frac{x_3 - x_1}{x_n - x_1}, for\ outlier\ x_1\ avoiding\ x_2,$$

The equations above are shown for the outlier detection of the lower values but can also be used for the higher values, meaning x_n and/or x_{n-1} [229]. Type10 (recommended for 3-7 samples) can only detect a single outlying window at the same genomic position over all samples, while type20 (recommended for 8-30 samples) can identify exactly two outlying windows, meaning the Q test will not detect outliers if more than 2 outliers are present. An example is illustrated in Figure 4.2. Depending on the sample size, our method can be applied using type10 and type20 independently. For type20, the method first identify if one outlier can be detected. If not, it assumes that it is masked by a second deviant value. Therefore, it detects both the outliers. Dixon's Q test was applied for each window at the same position over all samples to identify gains or losses considered as outliers (Figure 4.4). Outliers were regarded as significant with a p-value of less than or equal to 0.01. In general, the higher the p-value cutoff, the higher the number of detected outliers but also the number of false positives, meaning the p-value is a tuning parameter for sensitivity of our method.

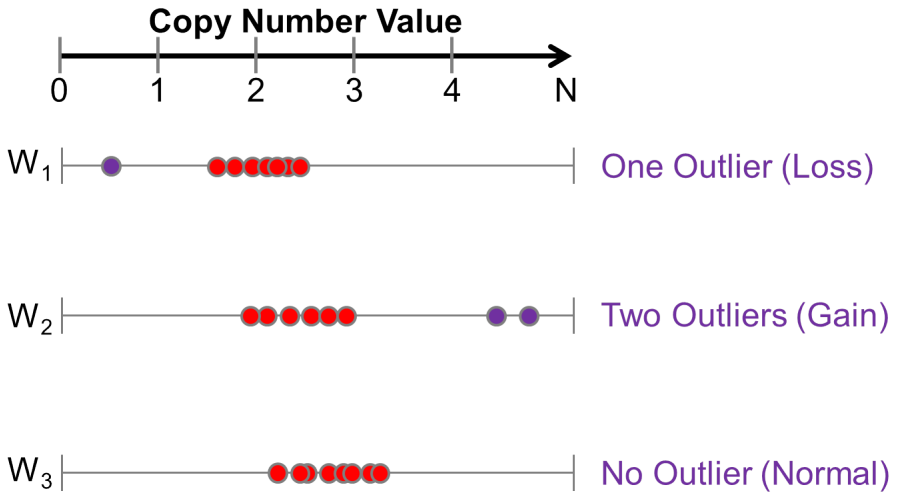


Figure 4.2: Different cases for Dixon's Q test when comparing Copy Number Values between different individuals. Here, three different windows are shown. Each dot represents the Copy Number Value for different individual. A purple dot represents the Copy Number Value of an individual considered as significant outlier (Loss or Gain). In W_1, the outlier is considered as loss because the value is less than all other individuals (red dots). In W_2, two outliers considered as gain and in W_3, there is no significant outlier.

4.2.3 Assessment of Outliers Using Hidden Markov Model

In the third and final step, the samples were again considered separately. For each sample, a Hidden Markov Model (HMM) was applied to get the most likely state of each window (i.e., gain, loss or normal). In general, HMM can be used to generate a sequence, that means to recover a series of states from a series of observation [230]. The parameters of a HMM are of two types, transition probabilities and emission probabilities. The transition probability is the probability of transitioning to a next state whereas emission probability is the probability of the observed variable emitted from a specific state (see Chapter 2.2.2.1). The initial transition and emission probabilities of the HMM are given in Figure 4.3. HMM is implemented to assess the outliers for each window. For example, if a given window is assigned as normal copy number and the previous windows shows copy number gain, it is likely that outlier was not detected in the second step. It is important to note that our method does not give the absolute copy number for a given region. This is due to the fact that targeted resequencing technology suffers with several biases such as local GC-content, as well as sequence complexity and sequence repetitiveness in the genome. Therefore, instead of absolute copy number, HMM is applied on the copy number state meaning

gain, loss or normal.

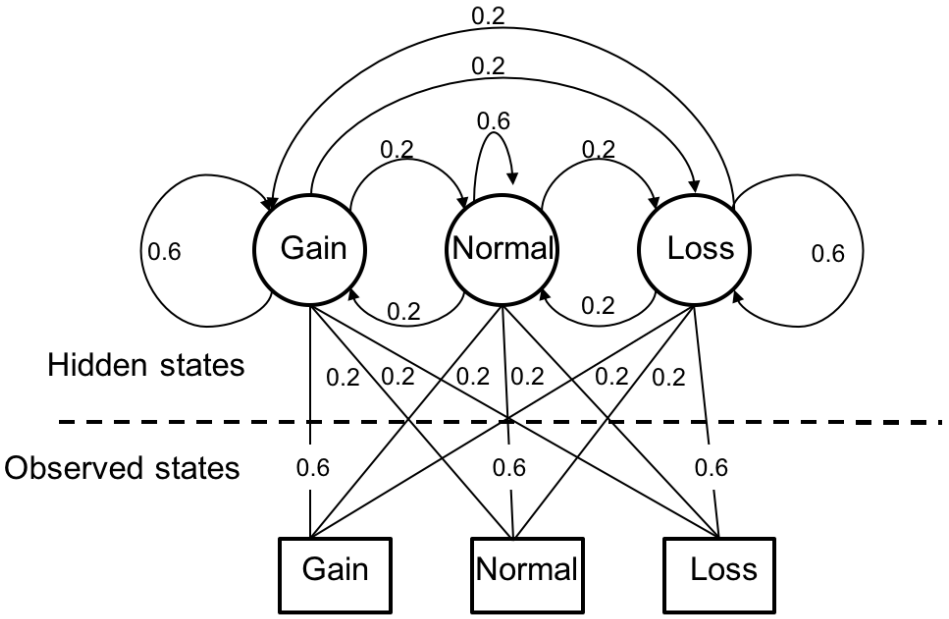


Figure 4.3: Initial transition and emission probabilities of the Hidden Markov Model. The transition probabilities are shown by curved arrows and emission probabilities are shown by lines.

The initial transition and emission probabilities of the HMM were recomputed using the Baum-Welch algorithm implemented in the R package 'HMM' (v1.0) [231]. The most likely sequence of the hidden states was then found by the Viterbi algorithm also implemented in the R package 'HMM' [232]. Finally, a region was called as copy number gain or loss if at least five continuous windows were considered as a gain or loss, respectively (Figure 4.4). This results in a minimum size of 500 bp for detectable CNVs.

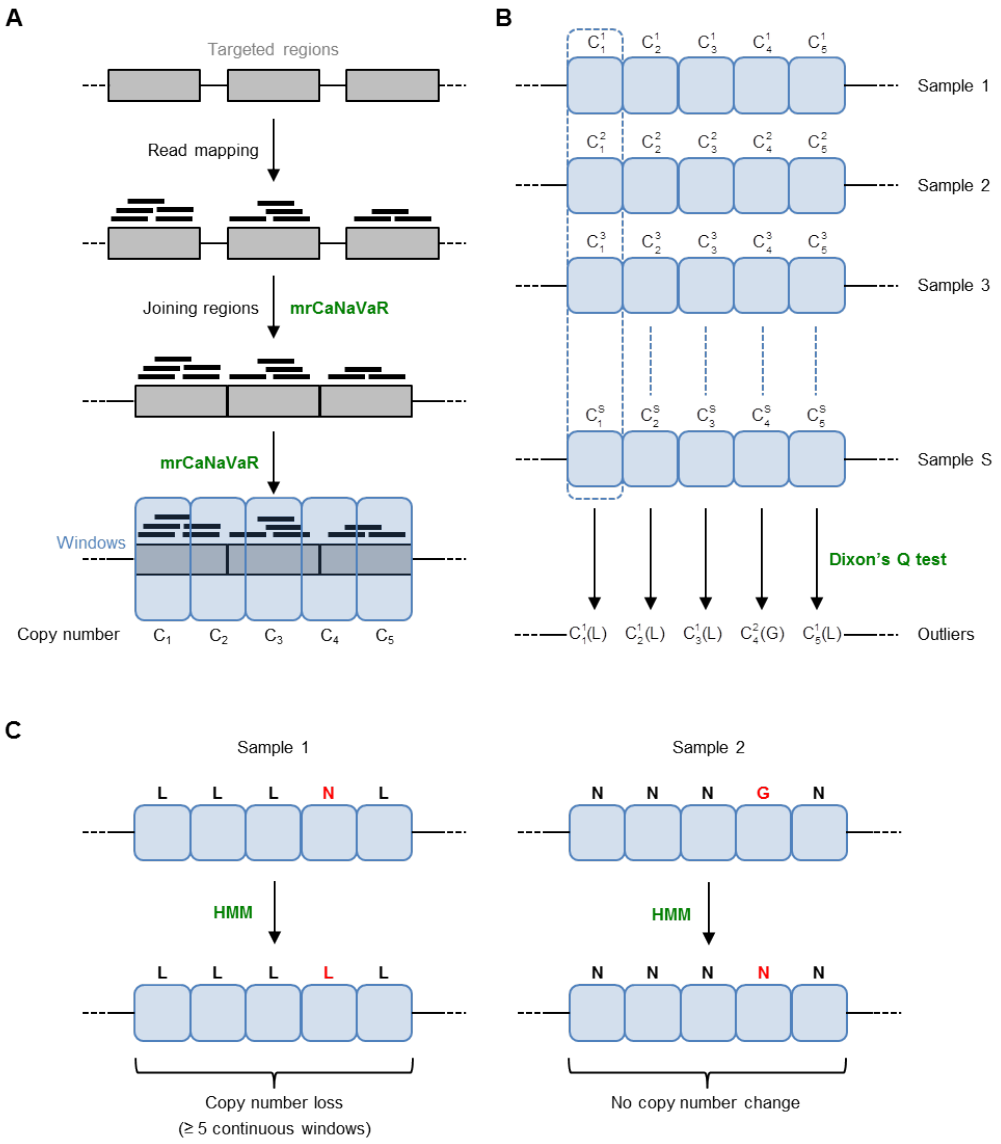


Figure 4.4: Outlier-based CNV calling method. (A) Read mapping and calculation of copy number value per window. Reads are mapped to extended targeted regions, which are then joined chromosome-wise. mrCaNaVaR is used to split the joined regions into windows. For each window, its copy number value is calculated by mrCaNaVaR, where C_W^S represents the value for window W in sample S. (B) Dixon's Q test is applied for each window over all samples to identify outliers. Here, sample 1 represents an outlier (loss, L) for the first, second, third and fifth window, while sample 2 represents an outlier (gain, G) for the fourth window. (C) Assessment of outliers using a Hidden Markov Model (HMM). In the given example, the fourth window of sample 1 is considered as normal (N). After applying the HMM, it will also be considered as a loss. Similarly, the fourth window of sample 2 is considered as normal after applying the HMM. A region is called as a copy number alteration, if at least five continuous windows show the same kind of change, i.e. either gain or loss.

4.3 Comparison of Outlier-Based Method

We applied our outlier-based CNV calling method to eight HapMap control samples and intersected our exome-based calls from five of the samples with previously generated calls from high-resolution microarray-based comparative genomic hybridization (array-CGH). In addition to our method, we used the two publicly available tools ExomeDepth and CoNIFER. Applying our method with type10 Dixon's Q test (assuming at most one outlier), we found 40 CNVs over the five HapMap controls (Table A.1), out of which 37 regions were also identified in the array-CGH data, showing a high positive predictive value of 93%. With type20 (assuming at most two outliers), we found 65 copy number changes (Appendix Table A.2), out of which 55 regions are present in the array-CGH data, resulting in a positive predictive value of 85%. Using CoNIFER, 32 CNVs were identified in the five HapMap exome controls and only 26 of these regions are also present in the array-CGH data [167], which corresponds to a positive predictive value of 81% (Appendix Table 4.1). Comparing our results to those obtained from CoNIFER, we found that with type10, 16 out of 40 regions (40%) are overlapping with regions called by CoNIFER by at least one base pair. Vice versa, 11 out of 32 regions (34%) overlap with our calls. With type20, 24 out of our 65 called regions (37%) overlap with those from CoNIFER and oppositely, 47% of the regions (15 out of 32) overlap with our calls. In addition to CoNIFER, we applied ExomeDepth with default parameters to the eight HapMap samples and intersected the found CNVs from five of the samples with previously generated calls from array-CGH. In summary, ExomeDepth found 1,555 CNVs in the five samples (median number of 286 CNVs per sample). Out of these, only 253 CNVs overlapped with 3,330 array-CGH calls, which suggest a positive predictive value of 16% and sensitivity of 7.6% (Table 4.1). Moreover, ExomeDepth identified more CNVs as compared to CoNIFER and to our method; however the positive predictive value is very low. Therefore, we decided not to use ExomeDepth for detecting CNVs in the TOF patients. To identify copy number alterations in TOF patients, we applied our outlier-based method as well as CoNIFER to targeted resequencing data of our eight cases. Using our method, we found four copy number gains in three genes, namely *ISL1*, *NOTCH1* and *PRODH*. CoNIFER only identified two gains in *PRODH*, which overlap with the two regions found by our method (Table 4.2).

Method	Number of CNVs	Validation dataset	Number of overlapping CNVs	Positive predictive value	Sensitivity
Outlier-based calling method with type10	40	3,330 arrayCGH calls	37	93%	1.1%
Outlier-based calling method with type20 including type10	65		55	85%	1.7%
CoNIFER	32		26	81%	0.8%
ExomeDepth	1,555		253	16%	7.6%

Table 4.1: Exome sequencing-based CNV calls in HapMap samples.

Method	Type of variation	Position (hg19)	Length in bp	Gene	Sample
Outlier-based calling method with type20 including type10	Gain	chr5:50,689,340-50,689,940	601	*ISL1*	TOF-23
	Gain	chr9:139,402,477-139,404,228	1,752	*NOTCH1*	TOF-01
	Gain	chr22:18,900,412-18,901,127	716	*PRODH*	TOF-02
	Gain	chr22:18,910,691-18,918,575	7,885	*PRODH*	TOF-02
CoNIFER	Gain	chr22:18,900,414-18,905,939	5,526	*PRODH*	TOF-02
	Gain	chr22:18,910,575-18,923,866	13,292	*PRODH*	TOF-02

Table 4.2: Targeted resequencing-based CNV calls in TOF patients.

4.4 Validation of Copy Number Variations

We further validated all four regions identified by our method using quantitative real-time PCR (Figure 4.5). Genomic DNA was extracted from whole blood or cardiac biopsies using standard procedures. Quantitative real-time PCR was carried out using GoTag qPCR Master Mix (Promega) on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems) according to the manufacturers instructions and with normalization to the *RPPH1* gene. As a reference, genomic DNA from the HapMap individual NA10851 was obtained from the Coriell Cell Repositories (New Jersey, USA).

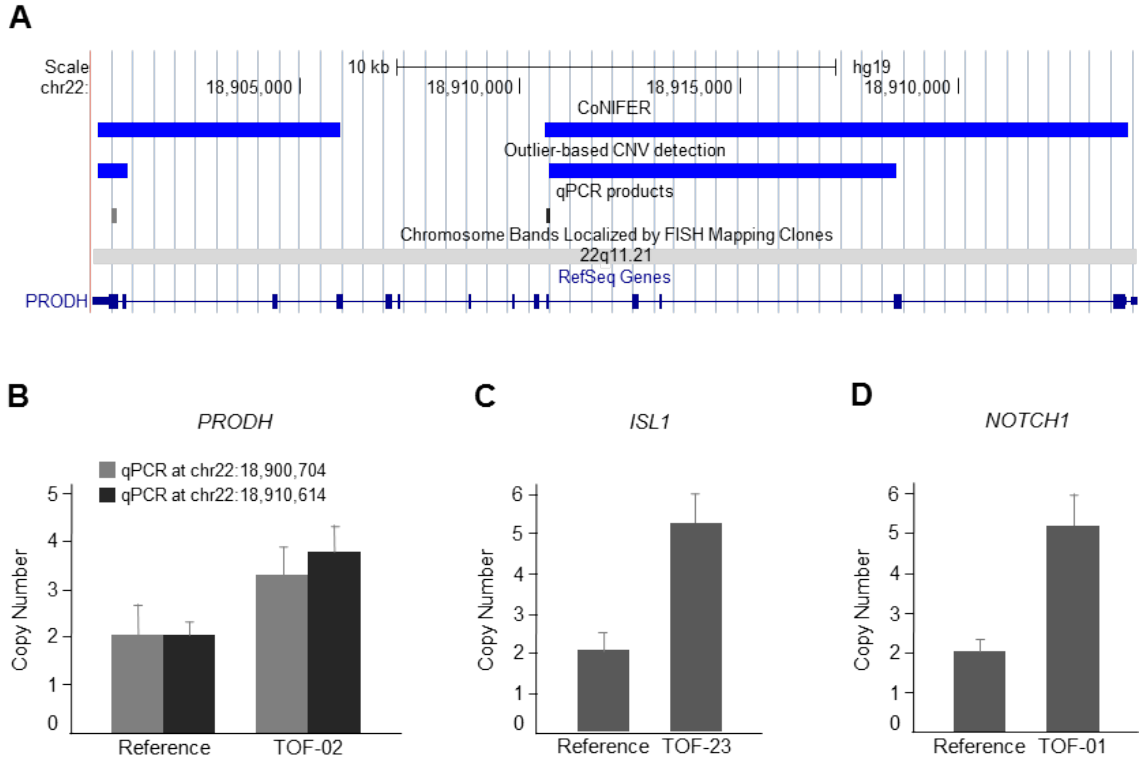


Figure 4.5: CNVs in TOF patients. (A) CNVs detected in *PRODH* by CoNIFER and our outlier-based CNV calling method. The duplications are depicted in the UCSC Genome Browser as blue bars. The positions of the two quantitative real-time PCR products selected for validation are shown as light and dark grey bars, respectively. (B) Quantitative real-time PCR validation of *PRODH* copy number gains. Measurement was performed at two different positions (light and dark grey bars, respectively) and normalized to the *RPPH1* gene. The HapMap individual NA10851 was used as a reference. The plot shows a representative of two independent measurements, which were each performed in triplicates. (C-D) Validation of copy number gains in *ISL1* and *NOTCH1*, respectively, that were only identified by our outlier-based CNV calling method.

4.5 Summary

In summary, we developed an outlier-based CNV calling method for a small cohort size of up to 30 individuals. Copy number variations (CNVs) are associated with a variety of diseases such as congenital heart defects and can be identified by high-throughput sequencing technologies. Our method is based on the assumption that individual CNVs (outliers) are disease-relevant and can be applied to exome as well as targeted resequencing data. Both sequencing techniques achieve a high read coverage over the targeted regions. Moreover, we assumed that the bias in the captured regions is similar in all samples enriched and sequenced with the same technology. We evaluated our method using publicly available data of eight HapMap samples and subsequently applied it to a small number of TOF patients. Compared to CoNIFER we identified more CNVs in both the HapMap samples as well as in our TOF cohort. In our TOF cohort comprising eight cases, we found four copy number gains in three patients, while CoNIFER only detected two of the gains in one patient. All four gains could be validated and in addition, the three genes affected by the CNVs are important regulators of heart development (*NOTCH1, ISL1*) or are located in a region associated with cardiac malformations (*PRODH*). Taken together, this illustrates the advantage of using an outlier-based detecting method in a small cohort with a heterogeneous genetic background. Thus, our method is of special interest for small cohorts of specific phenotypes like rare diseases. The method was implemented in R (v2.15.1) (Appendix Listing A.1).

Chapter 5

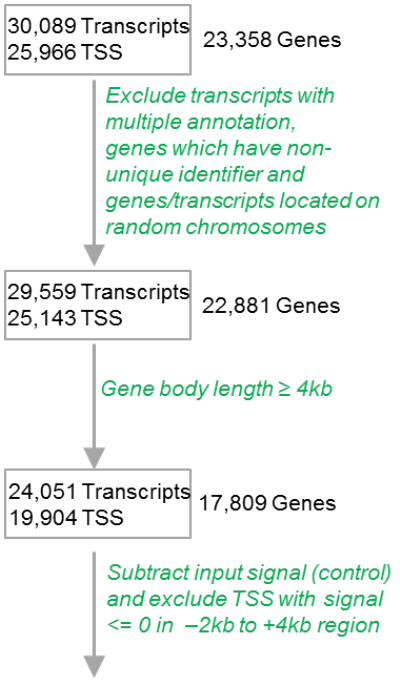
Analysis of Epigenetic Changes During Myogenic Differentiation of C2C12 Skeletal Muscle Cells

5.1 General Purpose

Myogenic differentiation is an essential process of muscle development and depends on the spatiotemporal regulation of gene expression patterns. Understanding myogenic differentiation is important to investigate muscular disease such as mucular dystrophies, which are regulated by epigenetic mechanisms [233, 234]. During myogenic differentiation, the overall content of histone methylations such as H3K4me2, H3K4me3, H3K36me3 and H3K27me3 were shown to be stable [81]. The basic helix-loop-helix (bHLH) transcription factor MyoD is a key player in myogenic specification and binds to DNA via a consensus E-box motif (CANNTG) [71, 72, 79]. During myogenesis, the binding of MyoD is primarily associated with gene activation [74], but its repressive function in myogenesis has also been shown on single genes [75–77]. Most previous studies focused on the dynamic regulation of histone modifications and transcription factors; however, it is still an open question how stable enrichment patterns of histone modifications in combinations with tissue specific transcription factors (TFs) regulate myogenic differentiation. Here, we investigated a stable enrichment pattern of the histone marks H3K4me2 and H3K4me3 in combinations with muscle tissue-specific transcription factor MyoD during myogenic differentiation. For this project, we analyzed ChIP-seq data of MyoD, H3K4me2 and H3K4me3 in undifferentiated (Undiff) and differentiated (Diff) C2C12 cells (Chapter 3.3). Furthermore, to compare the genome-wide gene expression profile in the two stages, we analyzed the RNA-seq data (Chapter 3.4).

5.2 H3K4me2 Located Over Gene Body of Muscle Specific Genes

To analyze the distribution of histone modifications around transcriptional start sites (TSS), we filtered RefSeq (mm9) annotation file, for a defined set of gene transcripts longer than 4 kb, resulting in 24,051 transcripts with 19,904 unique TSS (Figure 5.1). We further analyzed regions from -2 kb to +4 kb around TSS, which enables the direct comparison of epigenetic profiles independent from the gene length [194]. For each C2C12 ChIP-seq sample, the transcripts with a lower signal around the TSS as compared to the input sample were discarded, which resulted in approximately 18,000 to 20,000 transcripts per sample (Figure 5.1). For each TSS, we generated the average ChIP-seq profile based on the normalized signal.



Sample	TSS	Transcripts	Genes	
H3K4me2	14,881	18,332	13,384	Undiff C2C12
H3K4me3	15,577	19,185	14,147	
H3K4me2	15,925	19,536	14,275	Diff C2C12
H3K4me3	15,252	18,813	13,876	

Figure 5.1: Flow chart shows the filtering criteria and results for RefSeq genes (mm9).

In undifferentiated C2C12 cells, the average profile of H3K4me2 showed a bimodal distribution and revealed the highest enrichment downstream of the TSS (Appendix Figure B.1). To check whether any specific set of genes show distinct enrichment within the gene body, we performed k-means clustering with six clusters using the filtered set of TSS (Figure 5.2A). The first five clusters are characterized by specific distributions of H3K4me2 around the TSS. Here, cluster 1 (representing 632 genes) and cluster 4 (representing 1,361 genes) are the most distinct groups. Cluster 1 is characterized by H3K4me2 positioned over the gene body, while cluster 4 shows a higher prevalence upstream of the TSS. In contrast, cluster 6 includes all TSS with a very low H3K4me2 signal (4,942 genes). Finally, we verified the results of k-means clustering by discriminant analysis and found that the observed clusters can indeed be clearly distinguished from each other (Figure 5.2B).

When we combined the ChIP-seq data with gene expression profiles obtained by RNA-seq from ENCODE (Chapter 3.4), we found that genes located in cluster 1 are significantly higher expressed compared to all other clusters (P-value < 0.001). The lowest expression was found for genes detected in cluster 6 (Figure 5.2C). Further, we performed a GO enrichment analysis within each cluster using the DAVID functional annotation tool [192]. In contrast to all other clusters, cluster 1 genes are significantly enriched for GO terms related to muscle development (Figure 5.2D).

Next, we performed the same analysis of H3K4me2 profiles and related gene expression for differentiated C2C12 cells and obtained results comparable to the undifferentiated cells (Appendix Figure B.2). Again, significant GO terms related to muscle development were observed for cluster 1 genes. Then, we compared the clusters between undifferentiated and differentiated C2C12 cells and found that a high proportion of genes remained in the same cluster after differentiation (Appendix Figure B.2). For example, 83% of cluster 1 genes show a stable profile, with the same H3K4me2 distribution in both undifferentiated and differentiated C2C12 cells. Most interestingly, GO analysis of the stable (overlap between Undiff and Diff) and dynamic (specific for Undiff or Diff) gene sets in cluster 1 revealed that only the stable gene set is significantly enriched for muscle-related GO terms.

5.3 H3K4me3 Located Towards the Gene Body of Muscle-specific Genes

Compared to H3K4me2, we observed a higher mean enrichment of H3K4me3 directly downstream of the TSS in both undifferentiated and differentiated C2C12 cells

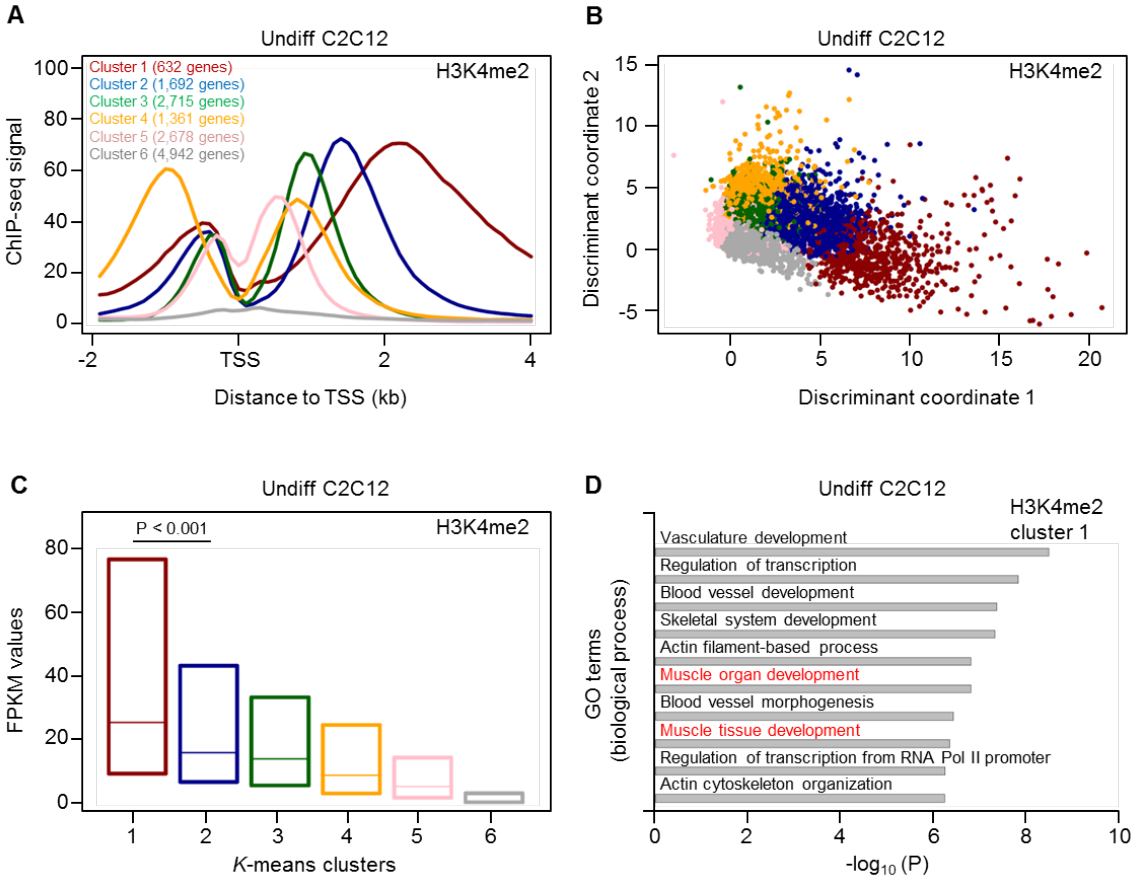


Figure 5.2: Clustering analysis of H3K4me2 profiles in undifferentiated C2C12 cells. (A) H3K4me2 profiles identified by k-means clustering. The clustering is based on TSS and the corresponding number of genes is given for each cluster. Genes with multiple TSS can be present in more than one cluster. (B) Discriminant analysis shows a clear distinction of the six clusters identified by k-means clustering. (C) The box plot (25% to 75% quartile) shows the levels of gene expression from the different H3K4me2 clusters in Undiff C2C12 cells. The expression of cluster 1 and cluster 2 genes was compared using the Mann-Whitney U test. (D) GO enrichment analysis of cluster 1 genes using the DAVID functional annotation tool. Top ten biological process terms with an adjusted (Benjamini-Hochberg) P-value ≤ 0.01 are indicated. GO terms related to muscle development are highlighted in red.

(Appendix Figure B.1). Moreover, clustering identified distinct profiles for both myoblasts and myotubes, with cluster 1 genes showing H3K4me3 enrichment towards the gene body (Figure 5.3A and Appendix Figure B.3). In addition, cluster 1 genes are significantly higher expressed as compared to the remaining clusters (P-value $<$ 0.001, Figure 5.3B and Appendix Figure B.3) and show a significant enrichment of GO terms related to muscle development. As for H3K4me2, we also compared the H3K4me3 clusters between both differentiation stages. Again, we found a high proportion of genes remaining in the same cluster. However, H3K4me3 profiles are in general more dynamic, meaning that more genes change their clusters during differentiation (Appendix Figure B.3). For example, only 71% of cluster 1 genes have a stable H3K4me3 profile, while 83% of cluster 1 genes have a stable H3K4me2 profile. As for H3K4me2, genes in cluster 1 with a stable H3K4me3 profile are significantly enriched for GO terms related to muscle development. Finally, the GO terms "muscle organ development" and "muscle tissue development" were further confirmed for cluster 1 in H3K4me2 as well as H3K4me3 in undifferentiated and differentiated C2C12 cells (P-values $< 10^{-7}$) using the MGI database. Moreover, we found a significant overlap of genes (P-value $< 2.2 \times 10^{-16}$, Figure 5.3C) and a significant enrichment of GO terms related to muscle development for these 347 common cluster 1 genes (Figure 5.3D). Comparable results were obtained in differentiated C2C12 cells (Appendix Figure B.4).

Figure 5.3E shows the H3K4me2/3 profiles for a subset of common cluster 1 genes. The transcription factor Six4 (SIX Homeobox 4) directly activates MyoD expression in gene regulatory networks [235, 236]. Mef2d (Myocyte Enhancer Factor 2D) is an early marker of the myogenic lineage and is required for skeletal muscle regeneration [237, 238]. Klf3 (Kruppel-Like Factor 3) synergizes with serum response factor on KLF binding sites to regulate muscle-specific gene expression [239]. The myogenic factor Tpm1 (Tropomyosin 1) is essential for myotube formation [240]. Acta2 (Actin, Alpha 2, Smooth Muscle, Aorta) and Myh9 (Myosin, Heavy Chain 9, Non-Muscle) belong to the actin and the myosin family of proteins, respectively, which are essential for muscle cell structure and mobility [241].

5.4 Genome-wide DNA Binding of MyoD

We performed ChIP-seq analysis in both C2C12 differentiation stages to identify genome-wide binding sites of the basic helix-loop-helix (bHLH) transcription factor MyoD, a known key factor for muscle development. For ChIP-seq data, we performed

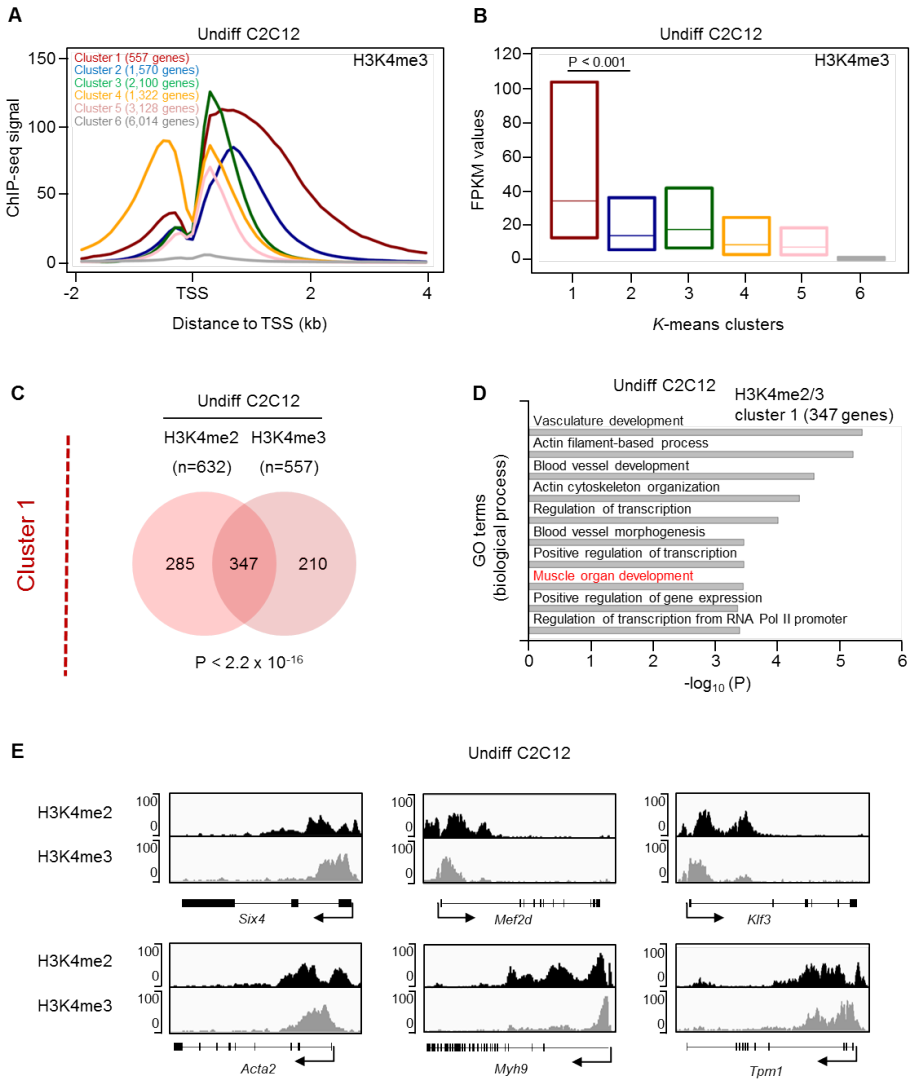


Figure 5.3: Clustering analysis of H3K4me3 profile in undifferentiated C2C12 cells and comparison to H3K4me2 profiles. (A) H3K4me3 profiles identified by k-means clustering. The clustering is based on TSS and the corresponding number of genes is given for each cluster. Genes with multiple TSS can be present in more than one cluster. (B) The box plot (25% to 75% quartile) shows the levels of gene expression from the different H3K4me3 clusters in Undiff C2C12 cells. The expression of cluster 1 and cluster 2 genes was compared using the Mann-Whitney U test. (C) Overlap of H3K4me2 and H3K4me3 cluster 1 genes in Undiff C2C12 cells. The P-value is based on a hypergeometric test. (D) GO enrichment analysis of common cluster 1 genes using the DAVID functional annotation tool. Top ten biological process terms with an adjusted (Benjamini-Hochberg) P-value ≤ 0.01 are indicated. GO terms related to muscle development are highlighted in red. (E) H3K4me2 and H3K4me3 enrichment profiles of selected muscle-relevant cluster 1 genes. The TSS is marked by an arrow. The y-axis indicates the ChIP-seq signal.

peak calling using MACS (v1.4.2) and assigned peaks to the genes if they are located within 10 kb upstream of the TSS or in the transcribed region. There were totally 6,069 and 22,934 ChIP-seq peaks, in undifferentiated and differentiated C2C12 cells, respectively. Genomic distribution of the peaks are shown in Figure 5.4. To confirm if the peaks are enriched for E-box motif, we used RSAT for the detection of novel motifs [188, 189]. Indeed the peaks were enriched for the expected motif (Figure 5.5). Furthermore, in-house (Sperling lab) script was generated to use TRANSFAC data and to confirm the enrichment of E-box motif (Appendix Listing B.1 and B.2). The in-house script uses the MATCH tool (Chapter 2.2.3.3), provided by TRANSFAC, for predicting transcription factor binding sites (TFBS) and consequently, identify statistically over-represented TFBS, assuming a binomial distribution. Using vertebrate non redundant profiles, we found the over-representation of "V$EBOX" matrix in both the stages.

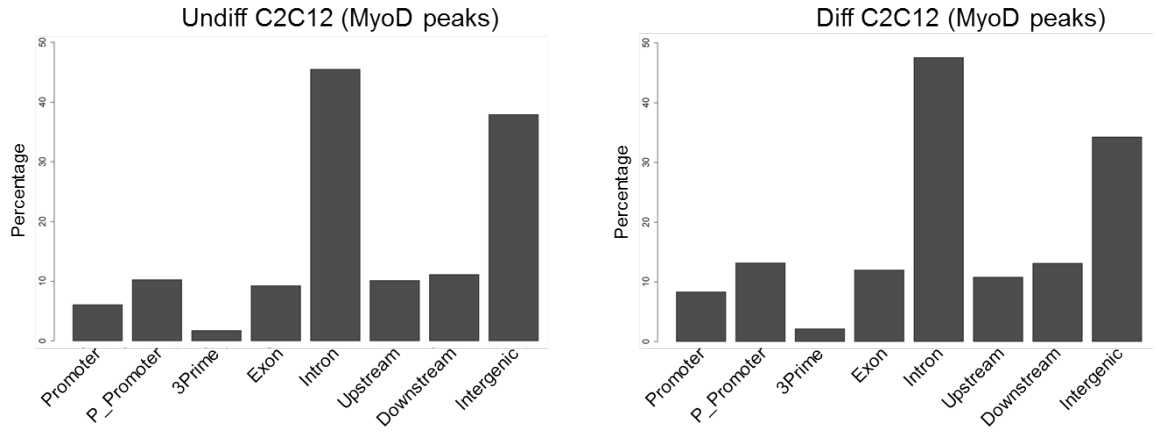


Figure 5.4: Genomic distribution of MyoD peaks in undifferentiated and differentiated C2C12 cells. P_Promoter, proximal promoter (Chapter 2.2.3.2).

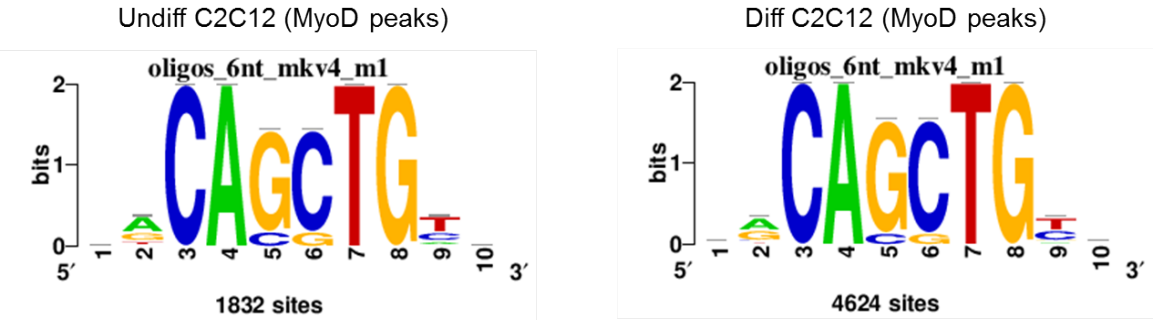


Figure 5.5: De novo motif analysis for MyoD peaks in undifferentiated and differentiated C2C12 cells.

MyoD ChIP-seq peaks were assigned to 2,813 and 7,477 genes in undifferentiated and differentiated C2C12 cells, respectively. To validate the success of our ChIP-seq experiment, we compared the data from ENCODE project with accession numbers ENCSR000AIG (undifferentiated C2C12 cells) and ENCSR000AIH (differentiated C2C12 cells). The significant overlap confirms that our data is comparable to the data provided by ENCODE (Figure 5.6).

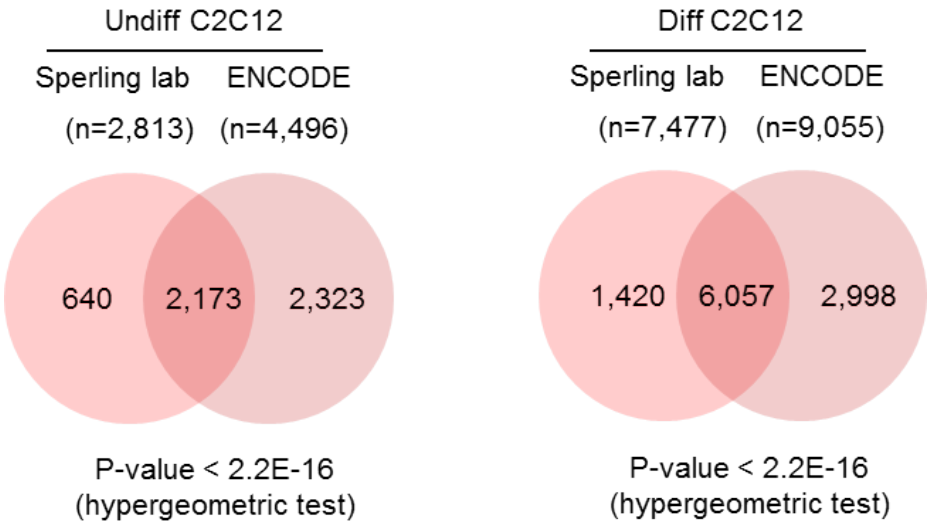


Figure 5.6: Comparison of our MyoD ChIP-seq data with ENCODE. Here numbers indicate the number of genes assigned.

5.5 Gene Expression During Myogenic Differentiation

RNA-seq data was used from the ENCODE project for undifferentiated C2C12 cells and differentiated C2C12 cells (60h timepoint). The data was downloaded from the Sequence Read Archive (SRA) at NCBI with accession numbers SRR496442 (undifferentiated C2C12 cells) and SRR496443 (differentiated C2C12 cells). RNA-seq reads were mapped to the mouse reference genome (mm9) using TopHat (v2.0.8) with default parameters. Furthermore, FPKM values were calculated using the Cufflinks (v2.0.2) with default parameters. To compare the genome-wide gene expression profile in the two stages, we plotted a scatter plot and found a very high correlation (Figure 5.7). This suggests that the gene expression of most of the genes is highly stable during C2C12 differentiation.

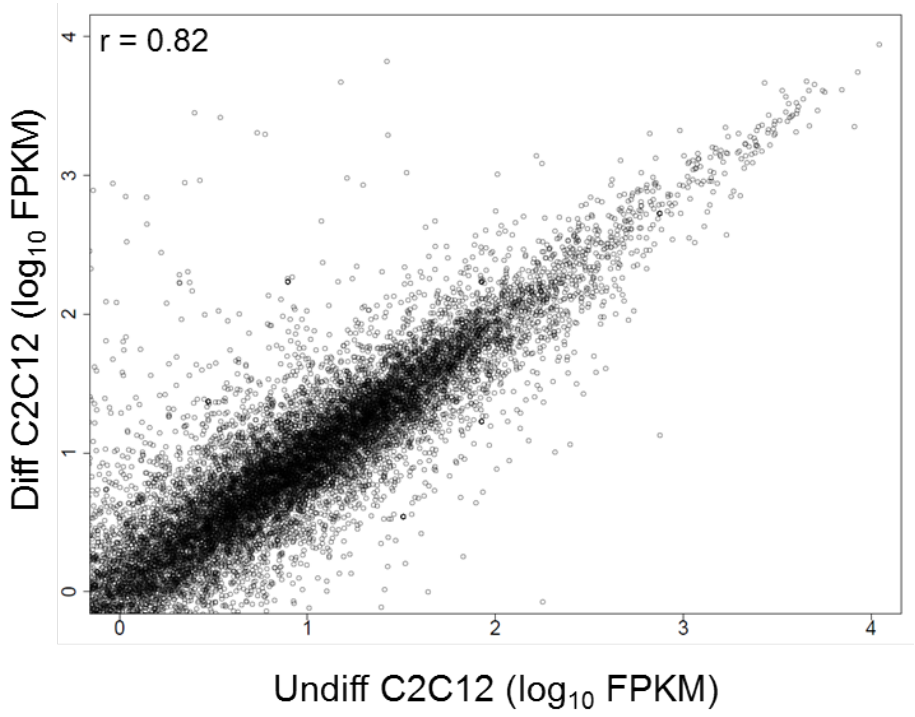


Figure 5.7: Comparison of the gene expression profile in undifferentiated and differentiated C2C12 cells. r is the Pearson's correlation.

We found that there are 10,584 genes with at least 1 FPKM in undifferentiated C2C12 cells and 11,015 genes with at least 1 FPKM in undifferentiated C2C12 cells. Genes with at least 1 FPKM in undifferentiated or differentiated C2C12 cells (11,381 genes) and with FC ≥ 2 were defined as differentially expressed genes. There were 1,698 genes up-regulated and 1,234 genes down-regulated during C2C12 differentiation. To confirm if the number of differentially expressed genes we found is comparable to the previous studies, we performed a literature search. Indeed, our results are in line with the previous studies (Table 5.1).

Reference	Up-regulated (genes)	Down-regulated (genes)	Experiment
Asp et .al., PNAS, 2011	1,620	Not given	Microarray
Rajan et. al., Physiological Genomics, 2012	~3,000	~3,500	Microarray
Hestand et. al, NAR, 2010	4,304		CAGE
Hestand et. al, NAR, 2010	3,846		SAGE

Table 5.1: Number of differentially expressed genes reported in the previous studies [81, 242, 243]. CAGE, Cap analysis gene expression; SAGE, Serial analysis of gene expression.

5.6 Expression of Cluster 1 Genes and Binding of MyoD

To identify muscle-relevant genes with a stable H3K4 di- and tri-methylation profile, we overlapped the 347 common cluster 1 genes from undifferentiated C2C12 cells with the 362 common cluster 1 genes from the differentiated stage. This resulted in a total of 267 genes with stable H3K4me2 and H3K4me3 profiles over or towards the gene body, respectively (Figure 5.8A). As expected, these 267 genes are significantly enriched for GO terms related to muscle development. Using RNA-seq data from ENCODE, 58 (22%) out of these genes are differentially expressed upon differentiation of C2C12 cells (fold change ≥ 2).

The percentage of genes bound by MyoD for each common H3K4me2/3 cluster is given in Figure 5.8B for both differentiation stages. Interestingly, cluster 1 harbors a significantly higher percentage of genes bound by MyoD compared to all other clusters (P-value < 0.01 in Undiff and P-value < 0.001 in Diff). We found approximately 30% of genes bound by MyoD in undifferentiated and 67% in differentiated C2C12 cells (Appendix Table B.1). Focusing again on the common stable cluster 1, out of the 267 genes, 95 gain MyoD during differentiation with 23% (22 genes) being differentially expressed (Figure 5.8A). As previous studies have mainly shown the activating role of MyoD [244, 245], we were expecting most of these genes to be up-regulated. Interestingly, we found 64% of this specific set of genes (14 out of 22) to be down-regulated, suggesting a repressive role of MyoD (Figure 5.8C).

In these 22 differentially expressed genes, we further searched for the MyoD binding E-box motif (CANNTG) within a 30bp region centered on the peak summit and found 14 genes harboring this motif (Figure 5.8C), of which 11 contain a particular E-box motif (CAGCTG) that has been shown to be preferred by MyoD during myogenic differentiation [79]. Among the down-regulated genes, five show the preferred E-box motif in their respective MyoD peaks (Figure 5.8C), including *Dusp6* (dual specificity phosphatase 6), *Frmd6* (FERM domain containing 6), *Patz1* (POZ (BTB) and AT hook containing zinc finger 1), *Ptbp1* (polypyrimidine tract binding protein 1) and Tmpo (thymopoietin). Most interestingly, the zinc finger transcription factor Patz1 was previously shown to have an important role in maintenance of the embryonic stem cell (ESC) phenotype and its knockdown leads to differentiation of murine ESCs into endoderm and mesoderm lineages at different time points [246].

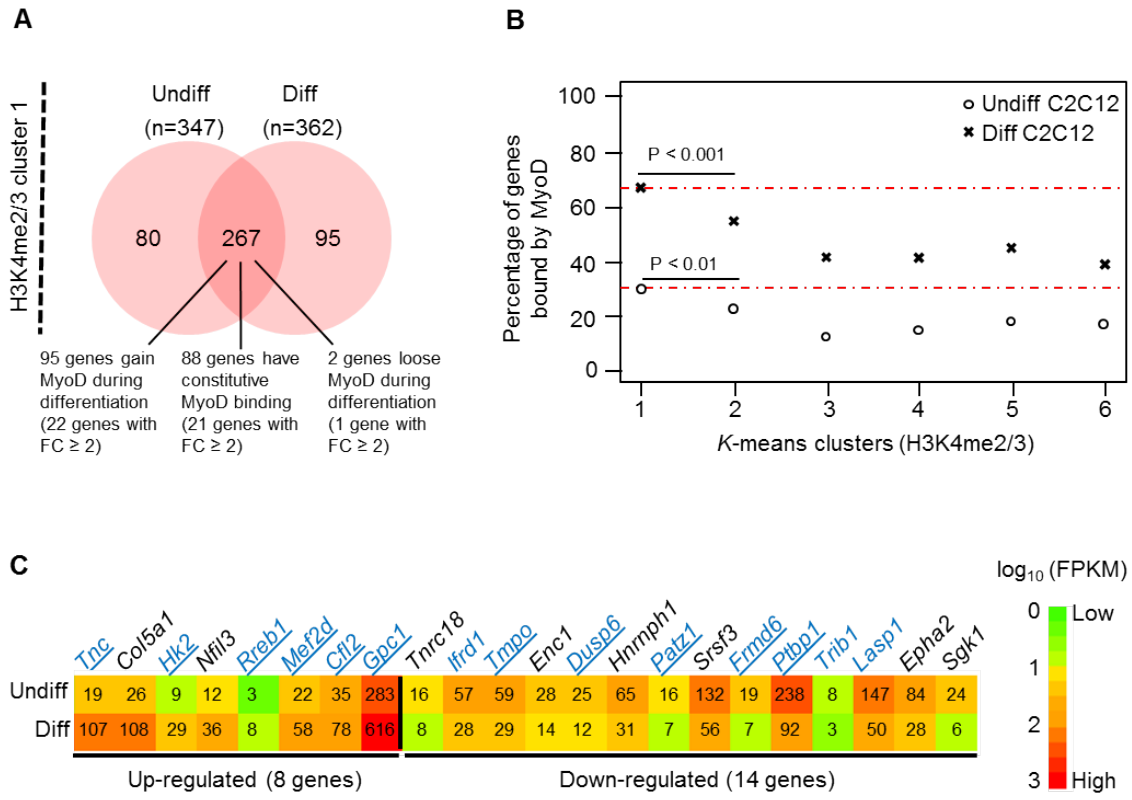


Figure 5.8: Cluster 1 genes bound by MyoD. (A) Overlap of common H3K4me2/3 cluster 1 genes in Undiff and Diff C2C12 cells. The number of genes, which gain, loose or have constitutive MyoD binding are indicated, including their respective number of differentially expressed genes (fold change (FC) ≥ 2). (B) Percentage of genes in the common H3K4me2/3 clusters bound by MyoD in Undiff and Diff C2C12 cells. Highest enrichment in Undiff and Diff (each in cluster 1) is indicated by the two red lines. The P-values are based on two-sided Fisher's exact test. (C) Heatmap of differentially expressed genes in H3K4me2/3 cluster1, which gain MyoD during differentiation (22 genes with fold change ≥ 2 out of 95 genes). The numbers in the heatmap represent the FPKM (fragments per kilo bases of exons for per million mapped) values. Gene names in blue indicate the genes with the E-box motif (CANNTG) within a 30 bp region centered on the peak summit. Gene names underlined in blue are genes with the MyoD preferred E-box motif (CAGCTG) within a 30 bp region centered on the peak summit.

5.7 Down regulation of *Patz1* by MyoD During Myogenic Differentiation

As shown before, *Patz1* shows a stable H3K4me2/3 profile and is bound by MyoD only in differentiated C2C12 cells (Figure 5.9A). Moreover, it is significantly down-regulated in the differentiated stage (Figure 5.9A). Both ChIP-seq and RNA-seq results were confirmed by real-time PCR (Figure 5.9B and Figure 5.9C, respectively). These results suggest that the binding of MyoD is related to the down-regulation of *Patz1* during differentiation of C2C12 cells.

To further investigate if the repression of *Patz1* is directly depending on MyoD expression during myogenic differentiation, IMR-90 human fibroblasts were converted to skeletal muscle cells by induction of MYOD. The expression of MYOD was induced ∼400-fold in growth medium (GM) and ∼800-fold in the differentiation medium (DM). Based on the ChIP-seq data, we found MYOD binding at the *PATZ1* promoter in both induced stages of converting fibroblasts to skeletal muscle cells (Figure 5.10A). Moreover, MYOD binding in the DM stage was higher compared to GM treated IMR-90 cells. Upon induction of MYOD, the expression of *PATZ1* was significantly down-regulated in both stages (Figure 5.10A, B). This finding could be confirmed by the induction of MYOD expression in BJ fibroblasts (Figure 5.10C).

Finally, we performed luciferase reporter gene assays to study the binding of MyoD at the *Patz1* promoter in vitro. Assays were performed by co-transfection of a MyoD expression vector together with a *Patz1* promoter reporter vector in HEK293 cells. The 400 bp core promoter of *Patz1* was efficiently repressed by co-transfection of MyoD in a dosage dependent manner.

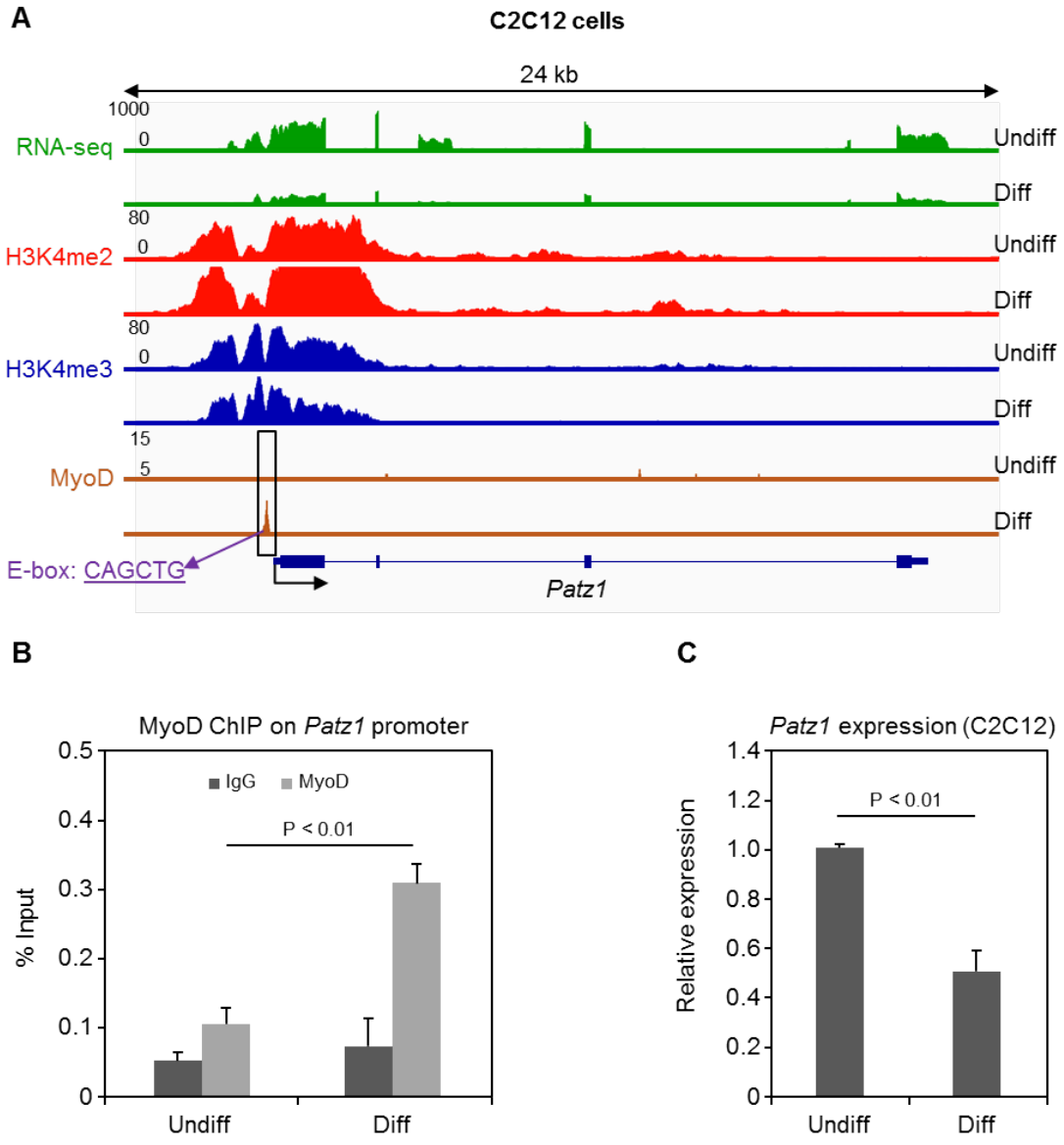


Figure 5.9: *Patz1* expression and MyoD binding during myogenic differentiation. (A) RNA expression profile of *Patz1*, H3K4me2 enrichment profile, H3K4me3 enrichment profile and MyoD binding profile at the *Patz1* promoter in Undiff and Diff C2C12 cells. All profiles are based on raw mapped reads. Position of the MyoD preferred E-box motif (CAGCTG) in the peak region is indicated. (B) ChIP analysis of MyoD occupancy levels at the *Patz1* promoter. Error bars indicate the standard deviation from at least three independent experiments. The statistical significance of enrichment versus the IgG control was calculated using Student's t-test. (C) Expression levels (mRNA) of *Patz1* in Undiff and Diff C2C12 cells were measured by real-time PCR in at least three independent experiments. The statistical significance of the difference in expression between Undiff and Diff C2C12 was calculated using Student's t-test.

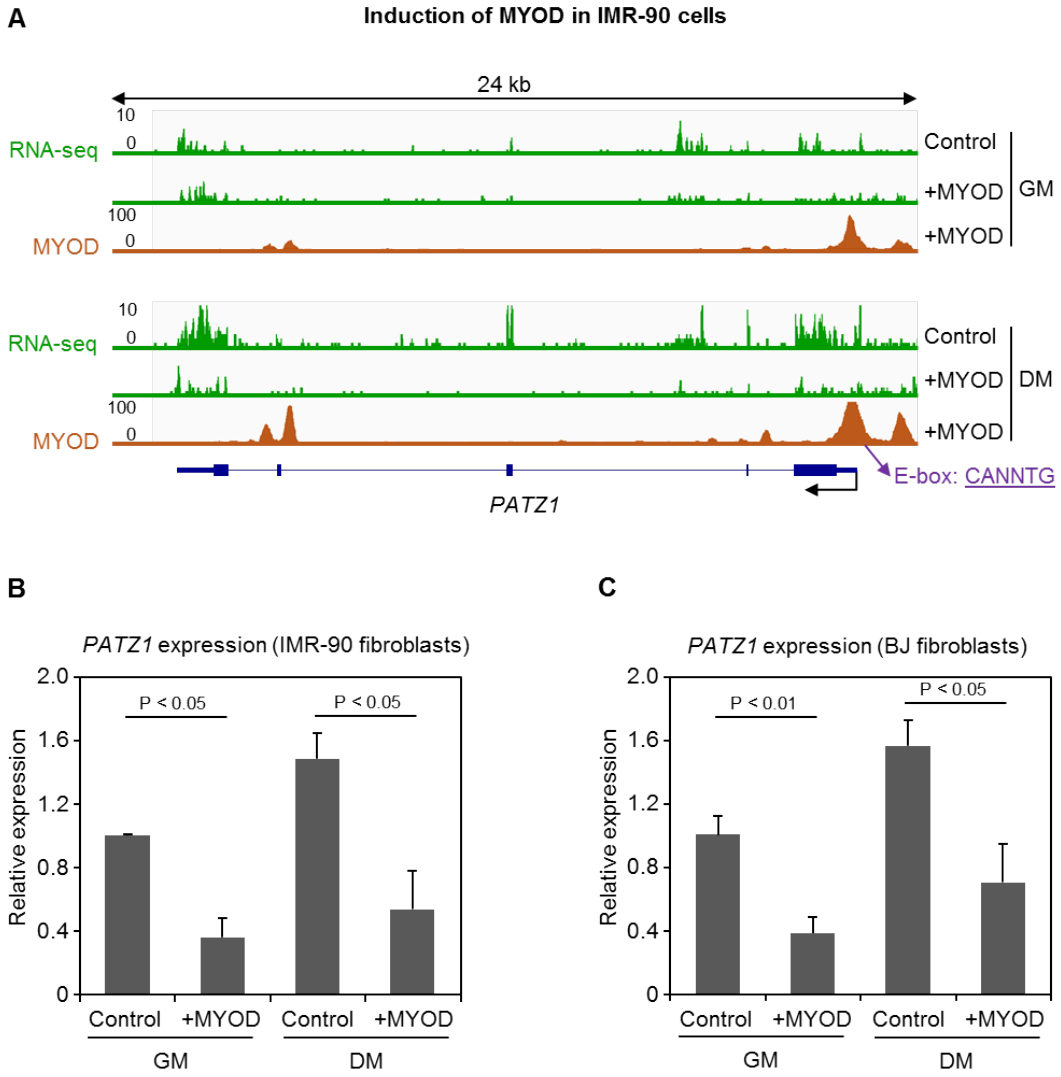


Figure 5.10: *PATZ1* down-regulation by MyoD. (A) RNA expression profile of *PATZ1* and MYOD binding profile at the *PATZ1* promoter in IMR-90 cells cultured with growth medium (GM) and differentiation medium (DM). All profiles are based on raw mapped reads. Position of the MyoD E-box motif (CANNTG) in the peak region is indicated. (B) Expression levels (mRNA) of *PATZ1* in IMR-90 fibroblasts. (C) Expression levels (mRNA) of *PATZ1* in BJ fibroblasts. P-value was calculated using Student's t-test based on at least three independent experiments.

5.8 Summary

In this study, we analyzed H3K4me2/3 signatures in myogenic differentiation and found specific profiles on muscle-relevant genes. In general, the average profile of H3K4me3 is enriched directly downstream of the TSS, whereas H3K4me2 is further located over the gene body, which has already demonstrated in hematopoietic cells [120, 247]. To identify specific H3K4me2/3 profiles, we used k-means clustering to define six groups of genes, showing distinct H3K4me2 and H3K4me3 patterns, respectively. We identified one cluster (cluster 1) with a H3K4 methylation profile over the gene body (di-methylation) or towards the gene body (tri-methylation), respectively. Moreover, the genes in cluster 1 are significantly higher expressed than all other clusters and are significantly enriched for GO terms related to muscle development. Furthermore, our study reveals a significantly higher binding of MyoD to this particular subset of genes and a predominantly repressive role of MyoD. It is important to note that MyoD is primarily associated with gene activation [74]. Our data also supports the activating role as most of genes that gain MyoD are up-regulated. Moreover, similar percentage of cluster 1 genes are differentially expressed, irrespective of gain of MyoD or constitutive MyoD binding during differentiation. Differential expression of the genes with constitutive MyoD binding could be explained by the regulation by other transcription factors. Interestingly most of the differentially expressed genes in the common stable cluster 1 that gain MyoD are down-regulated. Previously, repressive function of MyoD in myogenesis has also been shown on single genes [75–77]. Interestingly, further analysis and experiments revealed that MyoD binds and down-regulates Patz1 during myogenic differentiation. This observation was further confirmed in MyoD driven differentiation of fibroblasts to muscle cells. These findings might provide an important regulatory mechanism to promote myogenic differentiation.

Chapter 6

Detection of Differential Exon Usage in Dpf3 Knockout Mice

6.1 General Purpose

In previous studies, Sperling lab identified a chromatin remodeling factor Dpf3, whose expression was significantly up-regulated in the right ventricle of TOF patients [99, 100]. It was shown that the Dpf3 is specifically expressed in heart and somites and binds methylated and acetylated lysine residues of histone 3 and 4 [100]. Moreover, it is known that several histone modification-binding chromatin proteins interact with splicing factors [248]. Interestingly, experiments performed in Sperling lab showed that Dpf3 interacts with splicing factors which suggests its potential role in splicing. Therefore, to study the role of Dpf3 in splicing, Dpf3 knockout mice was generated. Using knockout (KO) and wildtype (WT) mice, we performed mRNA sequencing from 3 tissues, meaning right ventricle (RV), left ventricle (LV) and skeletal muscle (SM). In this study, we compared KO and WT mice to identify differential exon usage, meaning to identify the exons which are excluded or included due to Dpf3 knockout. We have generated a pipeline to analyse RNA-seq data for identifying differential exon usage without using replicates. For this project, we analyzed RNA-seq data from the right ventricle (RV), the left ventricle (LV) and the skeletal muscle (SM) of knockout (KO) and wildtype (WT) mice (Chapter 3.5).

6.2 Alignment of Reads to the Reference Sequence

The strand-specific libraries were prepared using "ScriptSeq RNA-seq library preparation kit" from Illumina and paired-end sequencing was performed on an Illumina HiSeq 2000 platform with the read length of 50 bp. The libraries were sequenced on

multiple lanes. RNA-seq reads were mapped to the mouse reference genome (mm9) using TopHat (v2.0.8) with the mate inner distance of 250 bp. The read realign edit distance was set to 0 for reporting the best possible alignment for the reads spanning multiple exons. Furthermore, coverage search and microexon search parameters were enabled to search for junctions and to find alignments incident to micro-exons, respectively. On average, 67% of the reads were mapped for right ventricle and left ventricle whereas 78% of the reads were mapped for skeletal muscle. Distribution of the reads is shown in Figure 6.1.

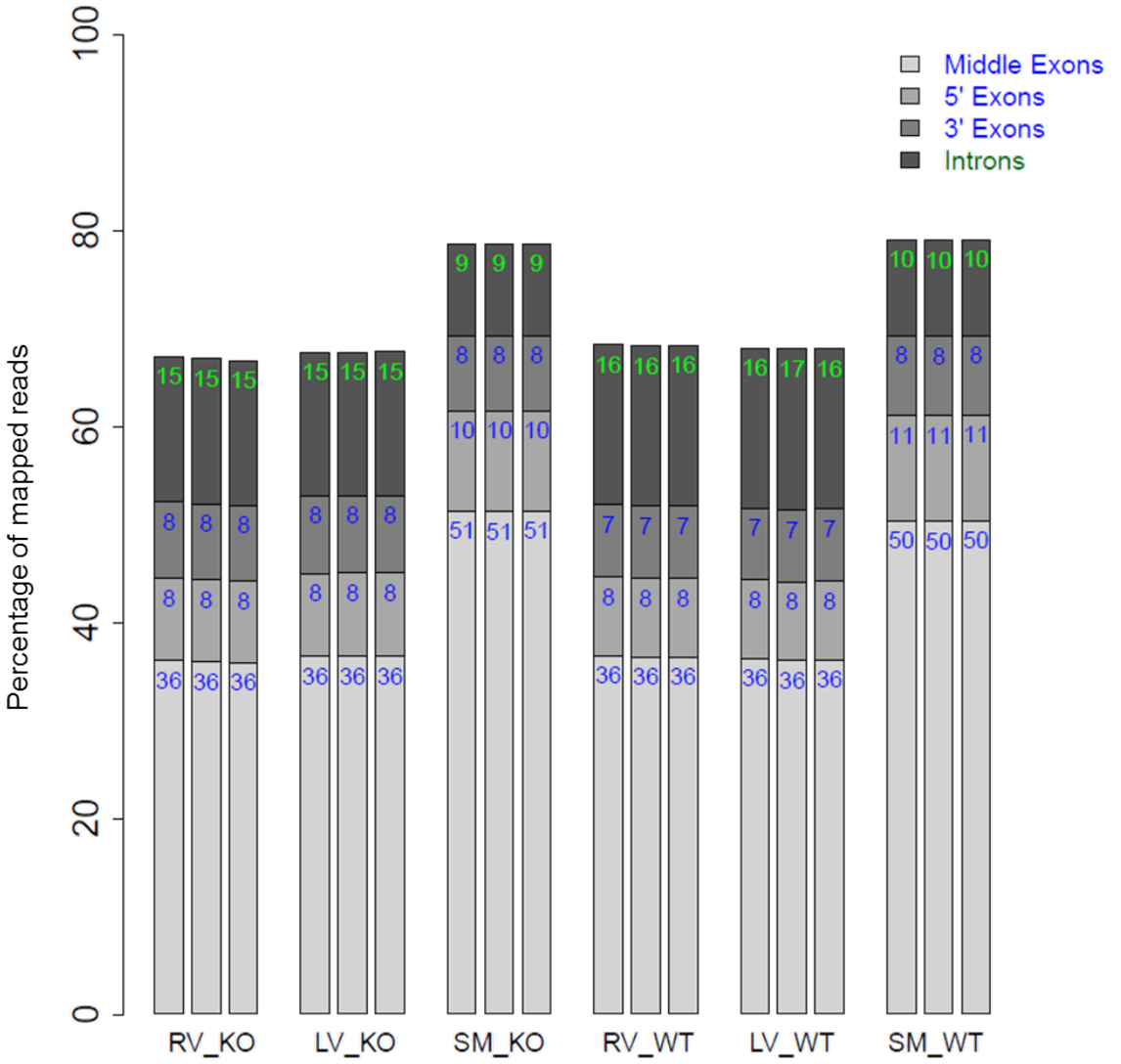


Figure 6.1: Distribution of RNA-seq reads. Mostly, reads are mapped to the middle exons. Numbers inside bars represents the percentage of reads. RV; right ventricle, LV; left ventricle, SM; skeletal muscle, KO; knockout, WT; wildtype.

In knockout mice, the second exon of the Dpf3 was deleted. Therefore, we checked the reads mapped on the second exon in WT and KO. As expected, there were no reads mapped to the second exon of Dpf3 in KO mice (Figure 6.2).

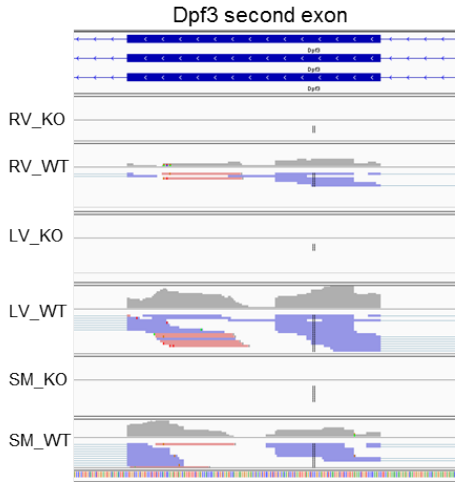


Figure 6.2: Reads mapped to the second exon of Dpf3. No reads were mapped in KO mice. RV; right ventricle, LV; left ventricle, SM; skeletal muscle, KO; knockout, WT; wildtype.

6.3 Computational Pipeline for Differential Exon Usage

The pipeline for the identification of the differential exon usage is based on the estimation of percent-spliced-in (PSI, Ψ) [218–220]. PSI values are calculated for each exon in both the conditions. Therefore, each exon will have two PSI values and difference between them (i.e. $\triangle$PSI) can be calculated. Before calculating $\triangle$PSI, we filtered for high confidence exons based on the minimum number of exonic and junction reads. Figure 6.3 illustrates the pipeline using the mapped files from right ventricle. In the first step, we calculate the fold change (FC) for all the mm9 exons. Next, we filtered out exons based on junction reads and exonic reads using the cutoff suggested by AltAnalyze [249]. In the last step, $\triangle$PSI is caluclated for each exon. Exclusion or inclusion of an exon is considered if $\triangle\text{PSI} \geq 10\%$. Results for all the samples are summarised in Table 6.1. One of the interesting example for right ventricle has been shown in Figure 6.4. *Myh7* encodes the beta heavy chain subunit of cardiac myosin and mutations in this gene are associated with hypertrophic cardiomyopathy and dilated cardiomyopathy [250, 251].

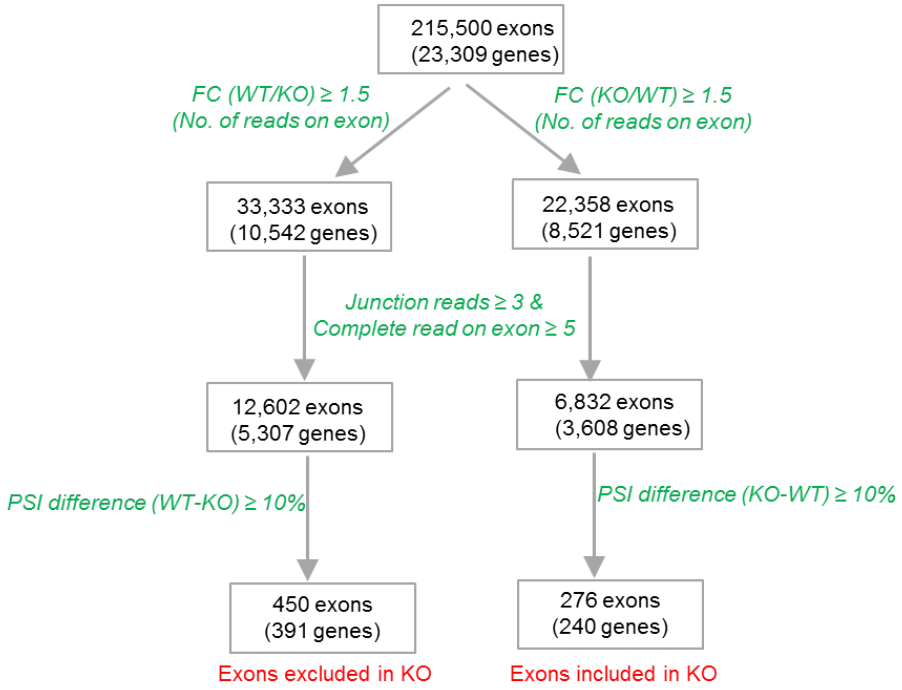


Figure 6.3: Pipeline for the identification of differential exon usage. The results are shown for the right ventricle. PSI; percent-spliced-in, KO; knockout, WT; wildtype.

Sample	Parameters	No. of exons Excluded (genes)	No. of exons Included (genes)	Example (excluded)
RV_KO vs RV_WT	FC ≥ 1.5 JR ≥ 3 & ER ≥ 5 PSI difference ≥ 10%	450 exons (391 genes)	276 exons (240 genes)	Exon2 of Dpf3, 9 exons of Myh7
LV_KO vs LV_WT	FC ≥ 1.5 JR ≥ 3 & ER ≥ 5 PSI difference ≥ 10%	507 exons (459 genes)	186 exons (173 genes)	Exon2 and 9 of Dpf3, 2 exons of Myh7
SM_KO vs SM_WT	FC ≥ 1.5 JR ≥ 3 & ER ≥ 5 PSI difference ≥ 10%	567 exons (473 genes)	186 exons (179 genes)	Exon2 of Dpf3, 3 exons of Myh7b

Table 6.1: Number of exons excluded or included in KO mice. RV; right ventricle, LV; left ventricle, SM; skeletal muscle, KO; knockout, WT; wildtype, FC; fold change, JR; junction reads, ER; exonic reads, PSI; percent-spliced-in.

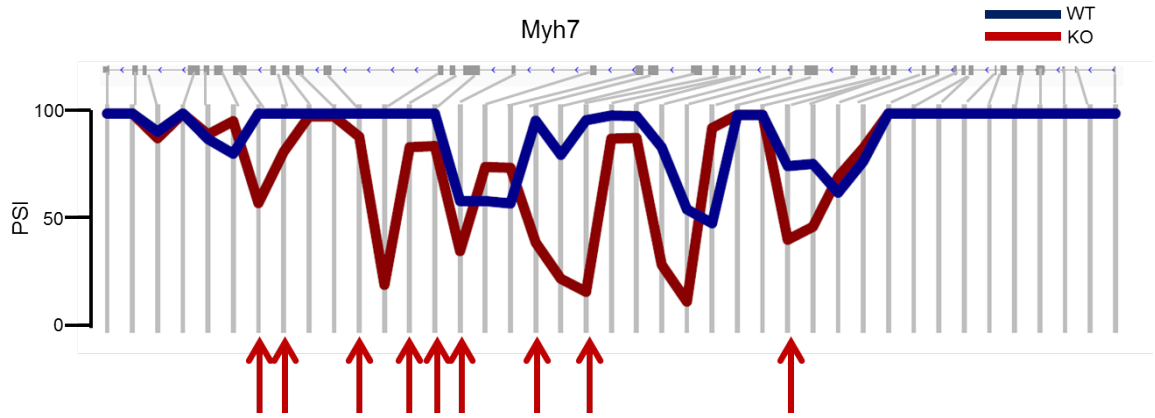


Figure 6.4: Exons of Myh7 excluded in Dpf3 KO mice. Excluded exons are indicated by red arrows. KO; knockout, WT; wildtype, PSI; percent-spliced-in.

6.4 Comparison to Alternative Splicing Detector

When we were generating the aforementioned pipeline, a tool was published to identify differential exon usage (DEU) without using replicates [216]. Therefore, we decided to apply this tool, ASD (Alternative Splicing Detector), on our dataset to compare the results to those from our pipeline. We used the same mapped files and the annotation files as we used in our pipeline. There were 43, 32 and 79 cassette exons found to be differentially used in right ventricle, left ventricle and skeletal muscle, respectively.

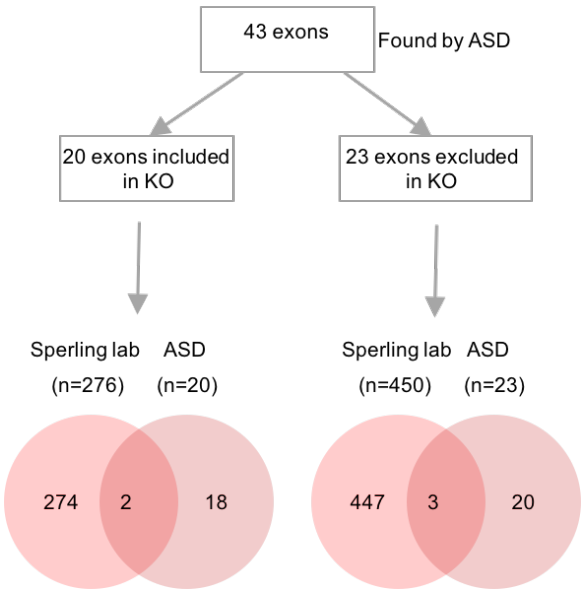


Figure 6.5: Comparison of Sperling lab pipeline and ASD for DEU in right ventricle. KO; knockout, ASD; Alternative Splicing Detector.

We found a very low overlap between the list generated by the Sperling lab pipeline and ASD. An example for right ventricle has been shown in Figure 6.5. The low overlap between the two pipelines suggests that the exons found are false positives or in other words, background. It is also evident in Figure 6.4 as most of the exons have very low PSI difference. Moreover, further lab experiments confirmed the same (data not shown).

6.5 Summary

In summary, we created a pipeline to identify differential exon usage from RNA-seq data without using replicates. To dissect the role of Dpf3 in splicing, knockout (KO) mice was generated and mRNA sequencing was performed from the right ventricle (RV), left ventricle (LV) and skeletal muscle (SM). We compared KO and WT mice to identify the exons which are excluded or included due to Dpf3 knockout. Moreover, our pipeline was compared with the published tool (i.e. ASD).

Chapter 7

Discussion

In this study, multiple high-throughput sequencing data have been analysed, namely targeted resequencing data from Tetralogy of Fallot (TOF) patients (Chapter 4), ChIP-seq and RNA-seq data during myogenic differentiation (Chapter 5) and RNA-seq data from Dpf3 knockout and wildtype mice (Chapter 6). The focus was on three different levels of gene regulation meaning genetic variations (Chapter 4), epigenetic regulation (Chapter 5) and splicing (Chapter 6).

In the first project (Chapter 4), we developed an outlier-based CNV calling method for a small cohort size of up to 30 individuals using targeted resequencing data. Copy number variations (CNVs) are associated with a variety of diseases such as congenital heart defects and can be identified by high-throughput sequencing technologies. In the past, microarray-based comparative genomic hybridization (array-CGH) and single nucleotide polymorphism (SNP) genotyping have been used commonly to resolve genomic changes. SNP arrays can be used to obtain the information of genotype as well as copy number changes but have limited ability to detect single-exon CNVs. On the other hand, features like high sequencing accuracy, low cost, coverage depth, experimental focus, and sample number make targeted resequencing more and more popular. Our CNV calling method is based on outlier detection using Dixon's Q test and outlier assessment using a Hidden Markov Model (HMM). We applied our outlier-based CNV calling method to eight HapMap control samples and intersected our exome-based calls from five of the samples with previously generated calls from high-resolution array-CGH. In addition to our method, we used the two publicly available tools ExomeDepth and CoNIFER [161, 167]. In comparison to CoNIFER, our method is able to detect more copy number alterations in small samples cohorts, demonstrated using exome data of eight HapMap samples as well as targeted resequencing data of eight Tetralogy of Fallot (TOF) samples. Furthermore, we applied ExomeDepth with default parameters to the eight HapMap samples and intersected the found

CNVs from five of the samples with previously generated calls from array-CGH. ExomeDepth identified more CNVs as compared to CoNIFER and to our method, however; the positive predictive value is very low. Therefore, we decided not to use ExomeDepth for detecting CNVs in the TOF patients.

As mentioned before, our method uses Dixon's Q test to identify outliers and subsequently, HMM to assess the outliers. In this study, Dixon's Q test is implemented in a way that it can detect outliers only in one direction for a particular window meaning either gain or loss. This makes sure that for a particular window, we do not consider many samples as outliers. For example, if we search for 2 outliers on each side when comparing a window between 8 samples, 50% of the samples will be considered as outliers. On the other hand, this decreases the true positive rate of our method. Nevertheless, our method can be extended to implement different types of Dixon's Q test. In addition to this, it is important to note that our method does not give the absolute copy number for a given region. This is due to the fact that targeted resequencing technology suffers with several biases such as local GC-content, as well as sequence complexity and sequence repetitiveness in the genome [168]. Therefore, instead of calculating absolute copy number, we can compare the same region across the samples to find out the copy number state meaning gain, loss or normal.

To identify copy number alterations in TOF patients, we applied our outlier-based method as well as CoNIFER to targeted resequencing data of our eight cases. Using our method, we found four copy number gains in three genes, namely *ISL1*, *NOTCH1* and *PRODH*. CoNIFER only identified two gains in *PRODH*, which overlap with the two regions found by our method. We further validated all four regions identified by our method using quantitative real-time PCR. *ISL1* is a homeobox transcription factor that marks cardiovascular progenitors and is known to be associated with human congenital heart disease [252, 253]. Interestingly, *Isl1* is required for the survival and migration of secondary heart field derived cells [254]. Moreover, it is known that the secondary heart field contributes to the development of right ventricle and the outflow tract (see Chapter 1). Considering the two phenotypes of TOF (i.e., right ventricular outflow tract obstruction and right ventricular hypertrophy), it highly suggests that the copy number gain of *ISL1* can lead to these phenotypes of TOF. *NOTCH1* is a transmembrane receptor involved in the NOTCH signaling pathway, which plays a crucial role in heart development [255]. Mutations in *NOTCH1* are associated with a spectrum of congenital aortic valve anomalies [256, 257] and a copy number loss was identified in a patient with TOF [55]. The mitochondrial protein *PRODH* catalyzes the first step in proline degradation and is located in the 22q11.2 locus. Deletions

in this region are associated with the DiGeorge syndrome and 80% of cases harbor cardiovascular anomalies [258]. A copy number gain and two losses in the 22q11.2 locus overlapping *PRODH* were also identified in sporadic TOF patients [55, 56].

The exploration of the human phenotype and its genetic and molecular background is the challenge of the next century and it is already clear that more precise phenotyping will lead to smaller cohort sizes. Moreover, analyzing small patient cohorts (3-30) is of special interest for rare diseases with only few available patient samples, for example trios. Approximately 7,000 rare diseases are currently known and together affect about 6% of the population [259]. Therefore, novel approaches like our CNV calling method, will be of exceptional relevance. Our method is based on the assumption that individual CNVs (outliers) are disease-relevant and can be applied to exome as well as targeted resequencing data.

In the second project (Chapter 5), we analyzed H3K4me2/3 signatures in myogenic differentiation and found specific profiles on muscle-relevant genes. In general, the average profile of H3K4me3 is enriched directly downstream of the TSS, whereas H3K4me2 is further located over the gene body, which has already demonstrated in hematopoietic cells [120, 247]. To identify specific H3K4me2/3 profiles, we used k-means clustering to define six groups of genes, showing distinct H3K4me2 and H3K4me3 patterns, respectively. K-means clustering has been shown to be useful in partitioning the distinct enrichment patterns of histone modifications [194, 195]. We identified one cluster (cluster 1) with a H3K4 methylation profile over the gene body (di-methylation) or towards the gene body (tri-methylation), respectively. Cluster 1 genes are significantly higher expressed than all other clusters and moreover, are significantly enriched for GO terms related to muscle development. This is in line with a previous study, showing a similar H3K4me2 profile of tissue-specific genes in CD4+ cells and brain tissue [194]. In addition, we could show that a unique profile (cluster 1) of H3K4me3 also marks muscle-specific genes.

Next, we overlapped the clusters identified in undifferentiated and differentiated C2C12 cells and found a high proportion of genes remaining in the same cluster for both methylation profiles, with H3K4me2 profiles being even more stable than H3K4me3 profiles. For acetylation of H3K9, H3K18 and H4K12, a striking reduction have previously described during myogenic differentiation, while di-methylation of H3K4 as well as tri-methylation of H3K4, H3K36 and H3K27 show more stable profiles [81]. We could further demonstrate that only the cluster 1 genes with stable profiles are enriched for GO terms related to muscle development. Moreover, we overlapped the common cluster 1 genes of H3K4me2/3 from undifferentiated C2C12 cells with the

common cluster 1 genes of H3K4me2/3 from the differentiated stage, which resulted in 267 genes with stable H3K4me2 and H3K4me3 profiles. Although methylation profiles are stable, we identified a considerable number of genes (58 out of 267 genes) differentially expressed upon differentiation of C2C12 cells, indicating an additional regulation by other factors. To further analyze the expression regulation of these common stable cluster 1 genes, we determined genome-wide binding of MyoD, which plays an essential role in activation of muscle-related genes [78–80].

MyoD typically binds promoters and enhancers of muscle-relevant genes to remodel the chromatin and activate transcription [74, 78, 260]. However, MyoD binding is not always associated with transcriptional activation [79]. For example, its repressive role was shown for the genes *Ccnb1* [75], *c-Fos* [76], and *Sp1* [77]. In our study, MyoD binding is significantly enriched in cluster 1 genes in undifferentiated and differentiated C2C12 cells. Interestingly, most of the differentially expressed genes with stable H3K4me2 and H3K4me3 profiles that gain MyoD in the myotubes are down-regulated, suggesting a repressive potential of MyoD. Five of the down-regulated genes harbor the preferred E-box motif in their MyoD peaks. The latter include *Dusp6*, which is a negative regulator of the MAP kinase superfamily and thus, plays a role in the regulation of proliferation and differentiation [261]. Moreover, *Dusp6* expression is also negatively regulated by the MyoD cofactor Mef2a in skeletal and cardiac muscle [262, 263]. Another down-regulated gene is *Ptbp1*, an antagonist of RBM4, which in turn activates the selection of skeletal muscle-specific exons in alpha-tropomyosin mRNA [264]. Finally, the zinc finger TF Patz1 is an important regulator of pluripotency by maintaining embryonic stem cells in an undifferentiated state [246], suggesting that Patz1 plays a similar role in C2C12 cells. Moreover, Patz1 interacts with p53 to target genes that are associated with cell differentiation and apoptosis [265]. It is located in the DiGeorge syndrome region on chromosome 22q12 and plays a critical role in the control of cell growth and embryonic development, which has been demonstrated by neural tube and cardiac outflow tract defects in Patz1 knockout mice [266].

Patz1 is ubiquitously expressed at early stages of embryonic development and becomes more restricted at later stages with almost no detectable expression in somites [266]. In contrast, MyoD shows an increased expression pattern in somites during embryonic development [267], indicating that Patz1 may represent a target which can be negatively regulated by MyoD during skeletal muscle development. Indeed, we show that *Patz1* expression is strongly reduced, associated with MyoD binding at the Patz1 promoter upon myogenic differentiation. However, it remains unclear how

Patz1 regulates myogenic differentiation and further studies are needed to investigate the underlying mechanisms.

Out of the differentially expressed common stable cluster 1 genes, 40% (23 genes) show a differential MyoD binding, while 36% (21 genes) show a constitutive MyoD binding. The differential expression of the latter could possibly be explained by other cofactors necessary for transcriptional regulation. For example, several studies have shown that MyoD can activate gene transcription in cooperation with other factors such as E-proteins and the chromatin remodeling factor Baf60c [71–73]. Nevertheless, most of the differentially expressed genes with stable H3K4me2 and H3K4me3 profiles that gain MyoD in the myotubes are down-regulated, which highly suggests a predominant repressive role of MyoD. Taken together, these findings might provide an important regulatory mechanism to promote myogenic differentiation.

In the third project (Chapter 6), we developed a pipeline to detect differential exon usage from RNA-seq data without using replicates. To study the role of Dpf3 in splicing, Dpf3 knockout mice was generated. Using knockout (KO) and wildtype (WT) mice, we performed mRNA sequencing from the right ventricle (RV), the left ventricle (LV) and the skeletal muscle (SM). In this project, we compared KO and WT mice to identify differential exon usage, meaning to identify the exons which are excluded or included due to Dpf3 knockout. The pipeline for the identification of the differential exon usage is based on the estimation of percent-spliced-in (PSI, Ψ). We focused on exon skipping alternative splicing (AS) event, which accounts for nearly 40% of AS events in higher eukaryotes [268–270]. In exon skipping event, an exon known as a cassette exon is spliced out of the transcript. When comparing the two conditions, for example wildtype and knockout mice, the skipped exon in the wildtype is considered as included exon in the knockout. We also compared the results from our pipeline with a tool, ASD, which is superior to our pipeline in a way that it also focus on other types of AS events such as alternative 3' splice site, alternative 5' splice site, alternative first exon, alternative last exon, mutually exclusive exons and intron retention [216]. The alternative splice site events meaning alternative 3' splice site and 5' splice site events occur when multiple splice sites are present at one end of the exon and accounts for 18.4% and 7.9% of all AS events in higher eukaryotes, respectively [268–270]. One of the rarest AS event in vertebrates and invertebrates is intron retention, in which specific introns remain unspliced in polyadenylated transcripts [268, 271, 272]. The remaining three complex AS events meaning alternative first exon, alternative last exon and mutually exclusive exons, are less frequent [268, 271, 273, 274]. Considering that exon skipping event is the

most frequent type of AS, we focus only on this event initially. In the future, it would be of great interest to perform the comprehensive analysis for different AS events.

As mentioned before, our pipeline does not consider the biological variability among the RNA-seq data. It has been shown that biological replicates are crucial for analyzing RNA-seq data [275]. Liu *et al.* showed that increasing the number of biological replicates consistently increases the power significantly, regardless of sequencing depth [275]. Interestingly, a recently published tool, rMATS (replicates multivariate analysis of transcript splicing), suggested the importance of biological replicates to detect alternative splicing changes from RNA-seq data [276]. rMATS uses a hierarchical framework to model percent-spliced-in (PSI, Ψ), which simultaneously accounts for estimation uncertainty in individual replicates and variability among replicates [276]. One of the commonly used tool which handle replicates is DEXSeq, that uses generalized linear models and test for the deviation of read counts on individual exons from the counts of the whole gene [217]. Given the importance of taking biological variation into account, it would be highly beneficial to use rMATS and/or DEXSeq whenever the replicates are available.

In this thesis high-throughput sequencing experiments have been extensively used, namely targeted resequencing, ChIP-seq and RNA-seq. Continuous efforts are being made by scientific community to update these experiments and consequently, the data analysis. For instance, ChIP-exonuclease (ChIP-exo) is a modified ChIP-seq approach, which involves chromatin immunoprecipitation (ChIP) combined with lambda exonuclease digestion followed by high-throughput sequencing [277–279]. ChIP-exo methodology allows for high resolution mapping of transcription factor DNA sites at nearly single nucleotide resolution. Subsequently, novel computational pipelines, for example ExoProfiler and GEM, were generated to analyze ChIP-exo data [280, 281]. Moreover, a recent study by Starick *et al.* showed the advantage of using ChIP-exo, which led them to explain the binding of glucocorticoid receptor on the regions devoid of its classical DNA recognition sequence [281]. Furthermore, Serandour *et al.* showed that ChIP-exo is also feasible in primary tissue such as mouse liver [282]. Although, ChIP-exo has been extensively used for mapping of transcription factor DNA sites, it would be of great interest to resolve the organization of individual histones on a genomic scale as shown by Rhee *et al.* [283].

The gene expression from RNA-seq data is calculated in the units of FPKM (fragments per kilobase per million mapped fragments). FPKM (or RPKM for single-end reads) is the most common normalisation method as it facilitates the comparison between genes within a sample and between the samples. However, Wagner *et al.*

showed that RPKM measure is inconsistent among samples and suggested a new unit of expression, TPM (transcripts per million), which is a slight modification of RPKM [284]. Given the FPKM/RPKM values, we can easily compute TPM [285]

$$TPM_i = (\frac{FPKM_i}{\sum_j FPKM_j})10^6.$$

However, it remains difficult to state clearly the benefits of TPM over FPKM or vice versa, specially when the RNA is extracted from a mixture of cells. Interestingly, single-cell sequencing-based technologies are becoming more common [286]. For example, single-cell RNA sequencing (scRNA-seq) can be used to perform accurate quantitative transcriptome measurement in individual cells [287]. A recently published strategy, Drop-seq, enables the transcriptional profiling of thousands of individual cells by encapsulating cells in tiny droplets for parallel analysis [288]. Drop-seq is a cost-effective strategy (preparation cost is around 2 to 5 cents per cell) for the analysis of heterogeneous mixture of cells from a tissue. Subsequently, an R package, Seurat, was designed for the analysis and visualization of scRNA-seq data. In future, it would be of great interest to apply this strategy on a tissue, like heart, to identify sub cell populations. However, it is important to note that it remains difficult to extract single cells from heart.

Data storage and accuracy of data analysis are the major challenges in the field of Bioinformatics. Although the price of storage devices is decreasing, the amount of data generated is increasing rapidly, which makes the data storage difficult even for large sequencing centers [289]. Furthermore, comparing and sharing this large amount of data is a challenging task. For example, the size of single sequenced human genome is approximately 140 gigabytes and comparing multiple genomes takes more than a personal computer [290]. It is difficult to share even a single sequenced genome using online file-sharing applications. The European Bioinformatics Institute (EBI) is one of the world's largest biology-data repositories, currently stores more than 20 petabytes of data [290]. Moreover, accurate analysis of this mass quantity of data poses an even larger challenge. Taken together, it is quiet evident that the current high-throughput technologies will be updated or might be replaced in future with more effective solutions. Therefore, with the advancement in the technology, more efforts will be required for the storage, sharing and comprehensive analysis of the data.

Bibliography

[1] Vikas Bansal et al. "Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with Tetralogy of Fallot." eng. In: *PLoS One* 9.1 (2014), e85375. DOI: 10.1371/journal.pone.0085375. URL: http://dx.doi.org/10.1371/journal.pone.0085375.

[2] Mark B. Gerstein et al. "What is a gene, post-ENCODE? History and updated definition." eng. In: *Genome Res* 17.6 (June 2007), pp. 669–681. DOI: 10.1101/gr.6339607. URL: http://dx.doi.org/10.1101/gr.6339607.

[3] G. Orphanides, T. Lagrange, and D. Reinberg. "The general transcription factors of RNA polymerase II." eng. In: *Genes Dev* 10.21 (Nov. 1996), pp. 2657–2683.

[4] Modan K. Das and Ho-Kwok Dai. "A survey of DNA motif finding algorithms." eng. In: *BMC Bioinformatics* 8 Suppl 7 (2007), S21. DOI: 10.1186/1471-2105-8-S7-S21. URL: http://dx.doi.org/10.1186/1471-2105-8-S7-S21.

[5] Lu Chen, Jaime M. Tovar-Corona, and Araxi O. Urrutia. "Alternative splicing: a potential source of functional innovation in the eukaryotic genome." eng. In: *Int J Evol Biol* 2012 (2012), p. 596274. DOI: 10.1155/2012/596274. URL: http://dx.doi.org/10.1155/2012/596274.

[6] Rudolf Jaenisch and Adrian Bird. "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals." eng. In: *Nat Genet* 33 Suppl (Mar. 2003), pp. 245–254. DOI: 10.1038/ng1089. URL: http://dx.doi.org/10.1038/ng1089.

[7] Bradley E. Bernstein, Alexander Meissner, and Eric S. Lander. "The mammalian epigenome." eng. In: *Cell* 128.4 (Feb. 2007), pp. 669–681. DOI: 10.1016/j.cell.2007.01.033. URL: http://dx.doi.org/10.1016/j.cell.2007.01.033.

[8] Tony Kouzarides. "Chromatin modifications and their function." eng. In: *Cell* 128.4 (Feb. 2007), pp. 693–705. DOI: 10.1016/j.cell.2007.02.005. URL: http://dx.doi.org/10.1016/j.cell.2007.02.005.

[9] Gabriel E. Zentner and Steven Henikoff. "Regulation of nucleosome dynamics by histone modifications." eng. In: *Nat Struct Mol Biol* 20.3 (Mar. 2013), pp. 259–266. DOI: 10.1038/nsmb.2470. URL: http://dx.doi.org/10.1038/nsmb.2470.

[10] Jenny J. Fischer et al. “Combinatorial effects of four histone modifications in transcription and differentiation.” eng. In: *Genomics* 91.1 (Jan. 2008), pp. 41–51. DOI: 10.1016/j.ygeno.2007.08.010. URL: http://dx.doi.org/10.1016/j.ygeno.2007.08.010.

[11] Raffaele Teperino, Kristina Schoonjans, and Johan Auwerx. “Histone methyl transferases and demethylases; can they link metabolism and transcription?” eng. In: *Cell Metab* 12.4 (Oct. 2010), pp. 321–327. DOI: 10.1016/j.cmet.2010.09.004. URL: http://dx.doi.org/10.1016/j.cmet.2010.09.004.

[12] Christoph Plass et al. “Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer.” eng. In: *Nat Rev Genet* 14.11 (Nov. 2013), pp. 765–780. DOI: 10.1038/nrg3554. URL: http://dx.doi.org/10.1038/nrg3554.

[13] Andrew J. Bannister and Tony Kouzarides. “Regulation of chromatin by histone modifications.” eng. In: *Cell Res* 21.3 (Mar. 2011), pp. 381–395. DOI: 10.1038/cr.2011.22. URL: http://dx.doi.org/10.1038/cr.2011.22.

[14] Robert Collins and Xiaodong Cheng. “A case study in cross-talk: the histone lysine methyltransferases G9a and GLP.” eng. In: *Nucleic Acids Res* 38.11 (June 2010), pp. 3503–3511. DOI: 10.1093/nar/gkq081. URL: http://dx.doi.org/10.1093/nar/gkq081.

[15] Lisa D. Moore, Thuc Le, and Guoping Fan. “DNA methylation and its basic function.” eng. In: *Neuropsychopharmacology* 38.1 (Jan. 2013), pp. 23–38. DOI: 10.1038/npp.2012.112. URL: http://dx.doi.org/10.1038/npp.2012.112.

[16] Zachary D. Smith and Alexander Meissner. “DNA methylation: roles in mammalian development.” eng. In: *Nat Rev Genet* 14.3 (Mar. 2013), pp. 204–220. DOI: 10.1038/nrg3354. URL: http://dx.doi.org/10.1038/nrg3354.

[17] Robert S. Illingworth and Adrian P. Bird. “CpG islands–’a rough guide’.” eng. In: *FEBS Lett* 583.11 (June 2009), pp. 1713–1720. DOI: 10.1016/j.febslet.2009.04.012. URL: http://dx.doi.org/10.1016/j.febslet.2009.04.012.

[18] T. Ordog et al. “Epigenetics and chromatin dynamics: a review and a paradigm for functional disorders”. In: *Neurogastroenterology and Motility* 24.12 (2012), pp. 1054–1068. ISSN: 1365-2982. DOI: 10.1111/nmo.12031. URL: http://dx.doi.org/10.1111/nmo.12031.

[19] Daniel Holoch and Danesh Moazed. “RNA-mediated epigenetic regulation of gene expression.” eng. In: *Nat Rev Genet* 16.2 (Feb. 2015), pp. 71–84. DOI: 10.1038/nrg3863. URL: http://dx.doi.org/10.1038/nrg3863.

[20] Thomas Brand. “Heart development: molecular insights into cardiac specification and early morphogenesis.” eng. In: *Dev Biol* 258.1 (June 2003), pp. 1–19.

[21] Marijke van Heeswijk, Jan G Nijhuis, and Hans MG Hollanders. “Fetal heart rate in early pregnancy”. In: *Early human development* 22.3 (1990), pp. 151–156.

[22] Jonathan A Epstein. "Cardiac development and implications for heart disease". In: *New England Journal of Medicine* 363.17 (2010), pp. 1638–1647.

[23] Marc Sylva, Maurice J B. van den Hoff, and Antoon F M. Moorman. "Development of the human heart." eng. In: *Am J Med Genet A* 164A.6 (June 2014), pp. 1347–1371. DOI: 10.1002/ajmg.a.35896. URL: http://dx.doi.org/10.1002/ajmg.a.35896.

[24] Mei Xin, Eric N. Olson, and Rhonda Bassel-Duby. "Mending broken hearts: cardiac development as a basis for adult heart regeneration and repair." eng. In: *Nat Rev Mol Cell Biol* 14.8 (Aug. 2013), pp. 529–541. DOI: 10.1038/nrm3619. URL: http://dx.doi.org/10.1038/nrm3619.

[25] Robert G. Kelly. "The second heart field." eng. In: *Curr Top Dev Biol* 100 (2012), pp. 33–65. DOI: 10.1016/B978-0-12-387786-4.00002-6. URL: http://dx.doi.org/10.1016/B978-0-12-387786-4.00002-6.

[26] Mariana Ruiz. *Tetralogy of Fallot*. June 2006.

[27] David J. McCulley and Brian L. Black. "Transcription factor pathways and congenital heart disease." eng. In: *Curr Top Dev Biol* 100 (2012), pp. 253–277. DOI: 10.1016/B978-0-12-387786-4.00008-7. URL: http://dx.doi.org/10.1016/B978-0-12-387786-4.00008-7.

[28] J. D. Molkentin et al. "Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis." eng. In: *Genes Dev* 11.8 (Apr. 1997), pp. 1061–1072.

[29] C. T. Kuo et al. "GATA4 transcription factor is required for ventral morphogenesis and heart tube formation." eng. In: *Genes Dev* 11.8 (Apr. 1997), pp. 1048–1060.

[30] I. Lyons et al. "Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5." eng. In: *Genes Dev* 9.13 (July 1995), pp. 1654–1666.

[31] Francisco J. Naya et al. "Mitochondrial deficiency and cardiac sudden death in mice lacking the MEF2A transcription factor." eng. In: *Nat Med* 8.11 (Nov. 2002), pp. 1303–1309. DOI: 10.1038/nm789. URL: http://dx.doi.org/10.1038/nm789.

[32] Zhiyv Niu et al. "Conditional mutagenesis of the murine serum response factor gene blocks cardiogenesis and the transcription of downstream gene targets." eng. In: *J Biol Chem* 280.37 (Sept. 2005), pp. 32531–32538. DOI: 10.1074/jbc.M501372200. URL: http://dx.doi.org/10.1074/jbc.M501372200.

[33] B. G. Bruneau et al. "A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease." eng. In: *Cell* 106.6 (Sept. 2001), pp. 709–721.

[34] Jun K. Takeuchi et al. “Tbx5 specifies the left/right ventricles and ventricular septum position during cardiogenesis.” eng. In: *Development* 130.24 (Dec. 2003), pp. 5953–5964. DOI: 10.1242/dev.00797. URL: http://dx.doi.org/10.1242/dev.00797.

[35] Chen-Leng Cai et al. “Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart.” eng. In: *Dev Cell* 5.6 (Dec. 2003), pp. 877–889.

[36] Joshua W. Vincentz et al. “Cooperative interaction of Nkx2.5 and Mef2c transcription factors during heart development.” eng. In: *Dev Dyn* 237.12 (Dec. 2008), pp. 3809–3819. DOI: 10.1002/dvdy.21803. URL: http://dx.doi.org/10.1002/dvdy.21803.

[37] Jenny Schlesinger et al. “The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs.” eng. In: *PLoS Genet* 7.2 (Feb. 2011), e1001313. DOI: 10.1371/journal.pgen.1001313. URL: http://dx.doi.org/10.1371/journal.pgen.1001313.

[38] Julien I E. Hoffman and Samuel Kaplan. “The incidence of congenital heart disease.” eng. In: *J Am Coll Cardiol* 39.12 (June 2002), pp. 1890–1900.

[39] Mark D. Reller et al. “Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005.” eng. In: *J Pediatr* 153.6 (Dec. 2008), pp. 807–813. DOI: 10.1016/j.jpeds.2008.05.059. URL: http://dx.doi.org/10.1016/j.jpeds.2008.05.059.

[40] Ashleigh A. Richards and Vidu Garg. “Genetics of congenital heart disease.” eng. In: *Curr Cardiol Rev* 6.2 (May 2010), pp. 91–97. URL: http://dx.doi.org/10.2174/157340310791162703.

[41] J. J. Schott et al. “Congenital heart disease caused by mutations in the transcription factor NKX2-5.” eng. In: *Science* 281.5373 (July 1998), pp. 108–111.

[42] Vidu Garg et al. “GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5.” eng. In: *Nature* 424.6947 (July 2003), pp. 443–447. DOI: 10.1038/nature01827. URL: http://dx.doi.org/10.1038/nature01827.

[43] Yung-Hao Ching et al. “Mutation in myosin heavy chain 6 causes atrial septal defect.” eng. In: *Nat Genet* 37.4 (Apr. 2005), pp. 423–428. DOI: 10.1038/ng1526. URL: http://dx.doi.org/10.1038/ng1526.

[44] JAMES J NORA. “Multifactorial inheritance hypothesis for the etiology of congenital heart diseases the genetic-environmental interaction”. In: *Circulation* 38.3 (1968), pp. 604–617.

[45] Robert B. Hinton. “Genetic and environmental factors contributing to cardiovascular malformation: a unified approach to risk.” eng. In: *J Am Heart Assoc* 2.3 (June 2013), e000292. DOI: 10.1161/JAHA.113.000292. URL: http://dx.doi.org/10.1161/JAHA.113.000292.

[46] Alan Fung et al. “Impact of prenatal risk factors on congenital heart disease in the current era.” eng. In: *J Am Heart Assoc* 2.3 (June 2013), e000064. DOI: 10.1161/JAHA.113.000064. URL: http://dx.doi.org/10.1161/JAHA.113.000064.

[47] Benoit G. Bruneau. “The developmental genetics of congenital heart disease.” eng. In: *Nature* 451.7181 (Feb. 2008), pp. 943–948. DOI: 10.1038/nature06801. URL: http://dx.doi.org/10.1038/nature06801.

[48] C. Ferencz et al. “Congenital heart disease: prevalence at livebirth. The Baltimore Washington Infant Study.” eng. In: *Am J Epidemiol* 121.1 (Jan. 1985), pp. 31–36.

[49] Christian Apitz, Gary D. Webb, and Andrew N. Redington. “Tetralogy of Fallot.” eng. In: *Lancet* 374.9699 (Oct. 2009), pp. 1462–1471. DOI: 10.1016/S0140-6736(09)60657-7. URL: http://dx.doi.org/10.1016/S0140-6736(09)60657-7.

[50] Akl C. Fahed et al. “Genetics of congenital heart disease: the glass half empty.” eng. In: *Circ Res* 112.4 (Feb. 2013), pp. 707–720. DOI: 10.1161/CIRCRESAHA.112.300853. URL: http://dx.doi.org/10.1161/CIRCRESAHA.112.300853.

[51] E. Goldmuntz et al. “Frequency of 22q11 deletions in patients with conotruncal defects.” eng. In: *J Am Coll Cardiol* 32.2 (Aug. 1998), pp. 492–498.

[52] Marcel Grunert et al. “Rare and private variations in neural crest, apoptosis and sarcomere genes define the polygenic background of isolated Tetralogy of Fallot.” eng. In: *Hum Mol Genet* 23.12 (June 2014), pp. 3115–3128. DOI: 10.1093/hmg/ddu021. URL: http://dx.doi.org/10.1093/hmg/ddu021.

[53] Silke R. Sperling. “Systems biology approaches to heart development and congenital heart disease.” eng. In: *Cardiovasc Res* 91.2 (July 2011), pp. 269–278. DOI: 10.1093/cvr/cvr126. URL: http://dx.doi.org/10.1093/cvr/cvr126.

[54] Kasper Lage et al. “Genetic and environmental risk factors in congenital heart disease functionally converge in protein networks driving heart development.” eng. In: *Proc Natl Acad Sci U S A* 109.35 (Aug. 2012), pp. 14035–14040. DOI: 10.1073/pnas.1210730109. URL: http://dx.doi.org/10.1073/pnas.1210730109.

[55] Steven C. Greenway et al. “De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot.” eng. In: *Nat Genet* 41.8 (Aug. 2009), pp. 931–935. DOI: 10.1038/ng.415. URL: http://dx.doi.org/10.1038/ng.415.

[56] Candice K. Silversides et al. “Rare copy number variations in adults with tetralogy of Fallot implicate novel risk gene pathways.” eng. In: *PLoS Genet* 8.8 (2012), e1002843. DOI: 10.1371/journal.pgen.1002843. URL: http://dx.doi.org/10.1371/journal.pgen.1002843.

[57] Rachel Soemedi et al. “Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease.” eng. In: *Am J Hum Genet* 91.3 (Sept. 2012), pp. 489–501. DOI: 10.1016/j.ajhg.2012.08.003. URL: http://dx.doi.org/10.1016/j.ajhg.2012.08.003.

[58] Adolfo Aguayo-Gomez et al. “Identification of Copy Number Variations in Isolated Tetralogy of Fallot.” eng. In: *Pediatr Cardiol* (June 2015). DOI: 10.1007/s00246-015-1210-9. URL: http://dx.doi.org/10.1007/s00246-015-1210-9.

[59] JL Krans. “The sliding filament theory of muscle contraction”. In: *Nature Education* 3.9 (2010), p. 66.

[60] A. F. HUXLEY and R. NIEDERGERKE. “Structural changes in muscle during contraction; interference microscopy of living muscle fibres.” eng. In: *Nature* 173.4412 (May 1954), pp. 971–973.

[61] H. HUXLEY and J. HANSON. “Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation.” eng. In: *Nature* 173.4412 (May 1954), pp. 973–976.

[62] C Florian Bentzinger, Yu Xin Wang, and Michael A. Rudnicki. “Building muscle: molecular regulation of myogenesis.” eng. In: *Cold Spring Harb Perspect Biol* 4.2 (Feb. 2012). DOI: 10.1101/cshperspect.a008342. URL: http://dx.doi.org/10.1101/cshperspect.a008342.

[63] B. Jostes, C. Walther, and P. Gruss. “The murine paired box gene, Pax7, is expressed specifically during the development of the nervous and muscular system.” eng. In: *Mech Dev* 33.1 (Dec. 1990), pp. 27–37.

[64] M. D. Goulding et al. “Pax-3, a novel murine DNA binding protein expressed during early neurogenesis.” eng. In: *EMBO J* 10.5 (May 1991), pp. 1135–1147.

[65] J. C. Kiefer and S. D. Hauschka. “Myf-5 is transiently expressed in nonmuscle mesoderm and exhibits dynamic regional changes within the presegmented mesoderm and somites I-IV.” eng. In: *Dev Biol* 232.1 (Apr. 2001), pp. 77–90. DOI: 10.1006/dbio.2000.0114. URL: http://dx.doi.org/10.1006/dbio.2000.0114.

[66] D. Sassoon et al. “Expression of two myogenic regulatory factors myogenin and MyoD1 during mouse embryogenesis.” eng. In: *Nature* 341.6240 (Sept. 1989), pp. 303–307. DOI: 10.1038/341303a0. URL: http://dx.doi.org/10.1038/341303a0.

[67] Y. Cinnamon et al. “The sub-lip domain–a distinct pathway for myotome precursors that demonstrate rostral-caudal migration.” eng. In: *Development* 128.3 (Feb. 2001), pp. 341–351.

[68] Mary Elizabeth Pownall, Marcus K. Gustafsson, and Charles P Emerson Jr. "Myogenic regulatory factors and the specification of muscle progenitors in vertebrate embryos." eng. In: *Annu Rev Cell Dev Biol* 18 (2002), pp. 747–783. DOI: 10.1146/annurev.cellbio.18.012502.105758. URL: http://dx.doi.org/10.1146/annurev.cellbio.18.012502.105758.

[69] Julianna Cseri. *Skeletal Muscle-from Myogenesis to Clinical Relations*. InTech, 2012.

[70] R. L. Davis, H. Weintraub, and A. B. Lassar. "Expression of a single transfected cDNA converts fibroblasts to myoblasts." eng. In: *Cell* 51.6 (Dec. 1987), pp. 987–1000.

[71] Shulamit Etzioni et al. "Homodimeric MyoD preferentially binds tetraplex structures of regulatory sequences of muscle-specific genes." eng. In: *J Biol Chem* 280.29 (July 2005), pp. 26805–26812. DOI: 10.1074/jbc.M500820200. URL: http://dx.doi.org/10.1074/jbc.M500820200.

[72] Stephen J. Tapscott. "The circuitry of a master switch: Myod and the regulation of skeletal muscle gene transcription." eng. In: *Development* 132.12 (June 2005), pp. 2685–2695. DOI: 10.1242/dev.01874. URL: http://dx.doi.org/10.1242/dev.01874.

[73] Sonia V. Forcales et al. "Signal-dependent incorporation of MyoD-BAF60c into Brg1-based SWI/SNF chromatin-remodelling complex." eng. In: *EMBO J* 31.2 (Jan. 2012), pp. 301–316. DOI: 10.1038/emboj.2011.391. URL: http://dx.doi.org/10.1038/emboj.2011.391.

[74] Yi Cao et al. "Global and gene-specific analyses show distinct roles for Myod and Myog at a common set of promoters." eng. In: *EMBO J* 25.3 (Feb. 2006), pp. 502–511. DOI: 10.1038/sj.emboj.7600958. URL: http://dx.doi.org/10.1038/sj.emboj.7600958.

[75] C. Chu, J. Cogswell, and D. S. Kohtz. "MyoD functions as a transcriptional repressor in proliferating myoblasts." eng. In: *J Biol Chem* 272.6 (Feb. 1997), pp. 3145–3148.

[76] D. Trouche et al. "Repression of c-fos promoter by MyoD on muscle cell differentiation." eng. In: *Nature* 363.6424 (May 1993), pp. 79–82. DOI: 10.1038/363079a0. URL: http://dx.doi.org/10.1038/363079a0.

[77] F. Vinals et al. "Myogenesis and MyoD down-regulate Sp1. A mechanism for the repression of GLUT1 during muscle cell differentiation." eng. In: *J Biol Chem* 272.20 (May 1997), pp. 12913–12921.

[78] Alexandre Blais et al. "An initial blueprint for myogenic differentiation." eng. In: *Genes Dev* 19.5 (Mar. 2005), pp. 553–569. DOI: 10.1101/gad.1281105. URL: http://dx.doi.org/10.1101/gad.1281105.

[79] Yi Cao et al. "Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming." eng. In: *Dev Cell* 18.4 (Apr. 2010), pp. 662–674. DOI: 10.1016/j.devcel.2010.02.014. URL: http://dx.doi.org/10.1016/j.devcel.2010.02.014.

[80] Roy Blum et al. "Genome-wide identification of enhancers in skeletal muscle: the role of MyoD1." eng. In: *Genes Dev* 26.24 (Dec. 2012), pp. 2763–2779. DOI: 10.1101/gad.200113.112. URL: http://dx.doi.org/10.1101/gad.200113.112.

[81] Patrik Asp et al. "Genome-wide remodeling of the epigenetic landscape during myogenic differentiation." eng. In: *Proc Natl Acad Sci U S A* 108.22 (May 2011), E149–E158. DOI: 10.1073/pnas.1102223108. URL: http://dx.doi.org/10.1073/pnas.1102223108.

[82] Giuseppina Caretti et al. "The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation." eng. In: *Genes Dev* 18.21 (Nov. 2004), pp. 2627–2638. DOI: 10.1101/gad.1241904. URL: http://dx.doi.org/10.1101/gad.1241904.

[83] Alexandre Blais et al. "Retinoblastoma tumor suppressor protein-dependent methylation of histone H3 lysine 27 is associated with irreversible cell cycle exit." eng. In: *J Cell Biol* 179.7 (Dec. 2007), pp. 1399–1412. DOI: 10.1083/jcb.200705051. URL: http://dx.doi.org/10.1083/jcb.200705051.

[84] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain terminating inhibitors." eng. In: *Proc Natl Acad Sci U S A* 74.12 (Dec. 1977), pp. 5463–5467.

[85] Martin Kircher and Janet Kelso. "High-throughput DNA sequencing–concepts and limitations." eng. In: *Bioessays* 32.6 (June 2010), pp. 524–536. DOI: 10.1002/bies.200900181. URL: http://dx.doi.org/10.1002/bies.200900181.

[86] Erwin L. van Dijk et al. "Ten years of next-generation sequencing technology." eng. In: *Trends Genet* 30.9 (Sept. 2014), pp. 418–426. DOI: 10.1016/j.tig.2014.07.001. URL: http://dx.doi.org/10.1016/j.tig.2014.07.001.

[87] International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome." eng. In: *Nature* 409.6822 (Feb. 2001), pp. 860–921.

[88] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome." eng. In: *Nature* 431.7011 (Oct. 2004), pp. 931–945.

[89] Jeffery A. Schloss. "How to get genomes at one ten-thousandth the cost." eng. In: *Nat Biotechnol* 26.10 (Oct. 2008), pp. 1113–1115. DOI: 10.1038/nbt1008-1113. URL: http://dx.doi.org/10.1038/nbt1008-1113.

[90] Lin Liu et al. "Comparison of next-generation sequencing systems." eng. In: *J Biomed Biotechnol* 2012 (2012), p. 251364. DOI: 10.1155/2012/251364. URL: http://dx.doi.org/10.1155/2012/251364.

[91] Cornelia Dorn, Marcel Grunert, and Silke R. Sperling. "Application of high-throughput sequencing for studying genomic variations in congenital heart disease." eng. In: *Brief Funct Genomics* 13.1 (Jan. 2014), pp. 51–65. DOI: 10.1093/bfgp/elt040. URL: http://dx.doi.org/10.1093/bfgp/elt040.

[92] Michael L. Metzker. "Sequencing technologies - the next generation." eng. In: *Nat Rev Genet* 11.1 (Jan. 2010), pp. 31–46. DOI: 10.1038/nrg2626. URL: http://dx.doi.org/10.1038/nrg2626.

[93] Jay Shendure et al. "Accurate multiplex polony sequencing of an evolved bacterial genome." eng. In: *Science* 309.5741 (Sept. 2005), pp. 1728–1732. DOI: 10.1126/science.1117389. URL: http://dx.doi.org/10.1126/science.1117389.

[94] Wilhelm J. Ansorge. "Next-generation DNA sequencing techniques." eng. In: *N Biotechnol* 25.4 (Apr. 2009), pp. 195–203. DOI: 10.1016/j.nbt.2008.12.009. URL: http://dx.doi.org/10.1016/j.nbt.2008.12.009.

[95] Elaine R. Mardis. "Next-generation DNA sequencing methods." eng. In: *Annu Rev Genomics Hum Genet* 9 (2008), pp. 387–402. DOI: 10.1146/annurev.genom.9.081307.164359. URL: http://dx.doi.org/10.1146/annurev.genom.9.081307.164359.

[96] Stephan C. Schuster. "Next-generation sequencing transforms today's biology." eng. In: *Nat Methods* 5.1 (Jan. 2008), pp. 16–18. DOI: 10.1038/nmeth1156. URL: http://dx.doi.org/10.1038/nmeth1156.

[97] Jay A. Shendure, Gregory J. Porreca, and George M. Church. "Overview of DNA sequencing strategies." eng. In: *Curr Protoc Mol Biol* Chapter 7 (Jan. 2008), Unit 7.1. DOI: 10.1002/0471142727.mb0701s81. URL: http://dx.doi.org/10.1002/0471142727.mb0701s81.

[98] John Eid et al. "Real-time DNA sequencing from single polymerase molecules." eng. In: *Science* 323.5910 (Jan. 2009), pp. 133–138. DOI: 10.1126/science.1162986. URL: http://dx.doi.org/10.1126/science.1162986.

[99] Bogac Kaynak et al. "Genome-wide array analysis of normal and malformed human hearts." eng. In: *Circulation* 107.19 (May 2003), pp. 2467–2474. DOI: 10.1161/01.CIR.0000066694.21510.E2. URL: http://dx.doi.org/10.1161/01.CIR.0000066694.21510.E2.

[100] Martin Lange et al. "Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex." eng. In: *Genes Dev* 22.17 (Sept. 2008), pp. 2370–2384. DOI: 10.1101/gad.471408. URL: http://dx.doi.org/10.1101/gad.471408.

[101] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. "The promise of whole-exome sequencing in medical genetics." eng. In: *J Hum Genet* 59.1 (Jan. 2014), pp. 5–15. DOI: 10.1038/jhg.2013.114. URL: http://dx.doi.org/10.1038/jhg.2013.114.

[102] Jacek Majewski et al. "What can exome sequencing do for you?" eng. In: *J Med Genet* 48.9 (Sept. 2011), pp. 580–589. DOI: 10.1136/jmedgenet-2011-100223. URL: http://dx.doi.org/10.1136/jmedgenet-2011-100223.

[103] Oxford Gene Technology. *NGS survey 2013*. April 2013.

[104] Lira Mamanova et al. "Target-enrichment strategies for next-generation sequencing." eng. In: *Nat Methods* 7.2 (Feb. 2010), pp. 111–118. DOI: 10.1038/nmeth.1419. URL: http://dx.doi.org/10.1038/nmeth.1419.

[105] Ryan Tewhey et al. "Microdroplet-based PCR enrichment for large-scale targeted sequencing." eng. In: *Nat Biotechnol* 27.11 (Nov. 2009), pp. 1025–1031. DOI: 10.1038/nbt.1583. URL: http://dx.doi.org/10.1038/nbt.1583.

[106] Gregory J. Porreca et al. "Multiplex amplification of large sets of human exons." eng. In: *Nat Methods* 4.11 (Nov. 2007), pp. 931–936. DOI: 10.1038/nmeth1110. URL: http://dx.doi.org/10.1038/nmeth1110.

[107] H. Johansson et al. "Targeted resequencing of candidate genes using selector probes." eng. In: *Nucleic Acids Res* 39.2 (Jan. 2011), e8. DOI: 10.1093/nar/gkq1005. URL: http://dx.doi.org/10.1093/nar/gkq1005.

[108] Emily Hodges et al. "Genome-wide in situ exon capture for selective resequencing." eng. In: *Nat Genet* 39.12 (Dec. 2007), pp. 1522–1527. DOI: 10.1038/ng.2007.42. URL: http://dx.doi.org/10.1038/ng.2007.42.

[109] Thomas J. Albert et al. "Direct selection of human genomic loci by microarray hybridization." eng. In: *Nat Methods* 4.11 (Nov. 2007), pp. 903–905. DOI: 10.1038/nmeth1111. URL: http://dx.doi.org/10.1038/nmeth1111.

[110] David T. Okou et al. "Microarray-based genomic selection for high-throughput resequencing." eng. In: *Nat Methods* 4.11 (Nov. 2007), pp. 907–909. DOI: 10.1038/nmeth1109. URL: http://dx.doi.org/10.1038/nmeth1109.

[111] Byung Yoon Choi et al. "Diagnostic application of targeted resequencing for familial nonsyndromic hearing loss." eng. In: *PLoS One* 8.8 (2013), e68692. DOI: 10.1371/journal.pone.0068692. URL: http://dx.doi.org/10.1371/journal.pone.0068692.

[112] *NimbleGen Sequence Capture Service*. http://www.atlas-biolabs.com/nimblegen.

[113] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." eng. In: *Nat Rev Genet* 10.1 (Jan. 2009), pp. 57–63. DOI: 10.1038/nrg2484. URL: http://dx.doi.org/10.1038/nrg2484.

[114] Ali Mortazavi et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." eng. In: *Nat Methods* 5.7 (July 2008), pp. 621–628. DOI: 10.1038/nmeth.1226. URL: http://dx.doi.org/10.1038/nmeth.1226.

[115] James Dominic Mills, Yoshihiro Kawahara, and Michael Janitz. "Strand-specific RNA-seq provides greater resolution of transcriptome profiling." eng. In: *Curr Genomics* 14.3 (May 2013), pp. 173–181. URL: http://dx.doi.org/10.2174/1389202911314030003.

[116] Joshua Z. Levin et al. "Comprehensive comparative analysis of strand-specific RNA sequencing methods." eng. In: *Nat Methods* 7.9 (Sept. 2010), pp. 709–715. DOI: 10.1038/nmeth.1491. URL: http://dx.doi.org/10.1038/nmeth.1491.

[117] *ScriptSeq RNA-Seq Library Preparation Kit*. http://www.illumina.com/products/scriptseq-rna-seq-library-prep.html.

[118] Terrence S. Furey. "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions." eng. In: *Nat Rev Genet* 13.12 (Dec. 2012), pp. 840–852. DOI: 10.1038/nrg3306. URL: http://dx.doi.org/10.1038/nrg3306.

[119] David S. Johnson et al. "Genome-wide mapping of in vivo protein-DNA interactions." eng. In: *Science* 316.5830 (June 2007), pp. 1497–1502. DOI: 10.1126/science.1141319. URL: http://dx.doi.org/10.1126/science.1141319.

[120] Artem Barski et al. "High-resolution profiling of histone methylations in the human genome." eng. In: *Cell* 129.4 (May 2007), pp. 823–837. DOI: 10.1016/j.cell.2007.05.009. URL: http://dx.doi.org/10.1016/j.cell.2007.05.009.

[121] Joshua W K. Ho et al. "ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis." eng. In: *BMC Genomics* 12 (2011), p. 134. DOI: 10.1186/1471-2164-12-134. URL: http://dx.doi.org/10.1186/1471-2164-12-134.

[122] Peter J. Park. "ChIP-seq: advantages and challenges of a maturing technology." eng. In: *Nat Rev Genet* 10.10 (Oct. 2009), pp. 669–680. DOI: 10.1038/nrg2641. URL: http://dx.doi.org/10.1038/nrg2641.

[123] Peggy J. Farnham. "Insights from genomic profiling of transcription factors." eng. In: *Nat Rev Genet* 10.9 (Sept. 2009), pp. 605–616. DOI: 10.1038/nrg2636. URL: http://dx.doi.org/10.1038/nrg2636.

[124] Axel Visel, Edward M. Rubin, and Len A. Pennacchio. "Genomic views of distant-acting enhancers." eng. In: *Nature* 461.7261 (Sept. 2009), pp. 199–205. DOI: 10.1038/nature08451. URL: http://dx.doi.org/10.1038/nature08451.

[125] Charles E. Massie and Ian G. Mills. "Mapping protein-DNA interactions using ChIP-sequencing." eng. In: *Methods Mol Biol* 809 (2012), pp. 157–173. DOI: 10.1007/978-1-61779-376-9_11. URL: http://dx.doi.org/10.1007/978-1-61779-376-9_11.

[126] Gregor D. Gilfillan et al. "Limitations and possibilities of low cell number ChIP-seq." eng. In: *BMC Genomics* 13 (2012), p. 645. DOI: 10.1186/1471-2164-13-645. URL: http://dx.doi.org/10.1186/1471-2164-13-645.

[127] Stephen G. Landt et al. "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." eng. In: *Genome Res* 22.9 (Sept. 2012), pp. 1813–1831. DOI: 10.1101/gr.136184.111. URL: http://dx.doi.org/10.1101/gr.136184.111.

[128] David Weese et al. "RazerS–fast read mapping with sensitivity control." eng. In: *Genome Res* 19.9 (Sept. 2009), pp. 1646–1654. DOI: 10.1101/gr.088823.108. URL: http://dx.doi.org/10.1101/gr.088823.108.

[129] Nuno A. Fonseca et al. "Tools for mapping high-throughput sequencing data." eng. In: *Bioinformatics* 28.24 (Dec. 2012), pp. 3169–3177. DOI: 10.1093/bioinformatics/bts605. URL: http://dx.doi.org/10.1093/bioinformatics/bts605.

[130] Matthew Ruffalo, Thomas LaFramboise, and Mehmet Koyutürk. "Comparative analysis of algorithms for next-generation sequencing read alignment." eng. In: *Bioinformatics* 27.20 (Oct. 2011), pp. 2790–2796. DOI: 10.1093/bioinformatics/btr477. URL: http://dx.doi.org/10.1093/bioinformatics/btr477.

[131] Ruiqiang Li et al. "SOAP: short oligonucleotide alignment program." eng. In: *Bioinformatics* 24.5 (Mar. 2008), p. 713. DOI: 10.1093/bioinformatics/btn025. URL: http://dx.doi.org/10.1093/bioinformatics/btn025.

[132] Heng Li, Jue Ruan, and Richard Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores." eng. In: *Genome Res* 18.11 (Nov. 2008), pp. 1851–1858. DOI: 10.1101/gr.078212.108. URL: http://dx.doi.org/10.1101/gr.078212.108.

[133] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." eng. In: *Genome Biol* 10.3 (2009), R25. DOI: 10.1186/gb-2009-10-3-r25. URL: http://dx.doi.org/10.1186/gb-2009-10-3-r25.

[134] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform." eng. In: *Bioinformatics* 25.14 (July 2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324. URL: http://dx.doi.org/10.1093/bioinformatics/btp324.

[135] Stephen M. Rumble et al. "SHRiMP: accurate mapping of short color-space reads." eng. In: *PLoS Comput Biol* 5.5 (May 2009), e1000386. DOI: 10.1371/journal.pcbi.1000386. URL: http://dx.doi.org/10.1371/journal.pcbi.1000386.

[136] Can Alkan et al. "Personalized copy number and segmental duplication maps using next-generation sequencing." eng. In: *Nat Genet* 41.10 (Oct. 2009), pp. 1061–1067. DOI: 10.1038/ng.437. URL: http://dx.doi.org/10.1038/ng.437.

[137] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq." eng. In: *Bioinformatics* 25.9 (May 2009), pp. 1105–1111. DOI: 10.1093/bioinformatics/btp120. URL: http://dx.doi.org/10.1093/bioinformatics/btp120.

[138] Kin Fai Au et al. "Detection of splice junctions from paired-end RNA-seq data by SpliceMap." eng. In: *Nucleic Acids Res* 38.14 (Aug. 2010), pp. 4570–4578. DOI: 10.1093/nar/gkq211. URL: http://dx.doi.org/10.1093/nar/gkq211.

[139] Kai Wang et al. "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery." eng. In: *Nucleic Acids Res* 38.18 (Oct. 2010), e178. DOI: 10.1093/nar/gkq622. URL: http://dx.doi.org/10.1093/nar/gkq622.

[140] Songbo Huang et al. "SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data." eng. In: *Front Genet* 2 (2011), p. 46. DOI: 10.3389/fgene.2011.00046. URL: http://dx.doi.org/10.3389/fgene.2011.00046.

[141] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner." eng. In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21. DOI: 10.1093/bioinformatics/bts635. URL: http://dx.doi.org/10.1093/bioinformatics/bts635.

[142] Paolo Ferragina and Giovanni Manzini. "Opportunistic data structures with applications". In: *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE. 2000, pp. 390–398.

[143] Michael Burrows and David J Wheeler. "A block-sorting lossless data compression algorithm". In: *Technical Report 124, Digital Equipment Corporation, Palo Alto, CA.* (1994).

[144] Ayat Hatem et al. "Benchmarking short sequence mapping tools." eng. In: *BMC Bioinformatics* 14 (2013), p. 184. DOI: 10.1186/1471-2105-14-184. URL: http://dx.doi.org/10.1186/1471-2105-14-184.

[145] Xiaoqing Yu et al. "How do alignment programs perform on sequencing data with varying qualities and from repetitive regions?" eng. In: *BioData Min* 5.1 (2012), p. 6. DOI: 10.1186/1756-0381-5-6. URL: http://dx.doi.org/10.1186/1756-0381-5-6.

[146] Heng Li and Richard Durbin. "Fast and accurate long-read alignment with Burrows-Wheeler transform." eng. In: *Bioinformatics* 26.5 (Mar. 2010), pp. 589–595. DOI: 10.1093/bioinformatics/btp698. URL: http://dx.doi.org/10.1093/bioinformatics/btp698.

[147] Ben Langmead and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." eng. In: *Nat Methods* 9.4 (Apr. 2012), pp. 357–359. DOI: 10.1038/nmeth.1923. URL: http://dx.doi.org/10.1038/nmeth.1923.

[148] Daehwan Kim et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." eng. In: *Genome Biol* 14.4 (2013), R36. DOI: 10.1186/gb-2013-14-4-r36. URL: http://dx.doi.org/10.1186/gb-2013-14-4-r36.

[149] Peter J. Campbell et al. "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing." eng. In: *Nat Genet* 40.6 (June 2008), pp. 722–729. DOI: 10.1038/ng.128. URL: http://dx.doi.org/10.1038/ng.128.

[150] Klaus Fellermann et al. "A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon." eng. In: *Am J Hum Genet* 79.3 (Sept. 2006), pp. 439–448. DOI: 10.1086/505915. URL: http://dx.doi.org/10.1086/505915.

[151] Bert B A. de Vries et al. "Diagnostic genome profiling in mental retardation." eng. In: *Am J Hum Genet* 77.4 (Oct. 2005), pp. 606–616. DOI: 10.1086/491719. URL: http://dx.doi.org/10.1086/491719.

[152] Bernard Thienpont et al. "Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients." eng. In: *Eur Heart J* 28.22 (Nov. 2007), pp. 2778–2784. DOI: 10.1093/eurheartj/ehl560. URL: http://dx.doi.org/10.1093/eurheartj/ehl560.

[153] F. Erdogan et al. "High frequency of submicroscopic genomic aberrations detected by tiling path array comparative genome hybridisation in patients with isolated congenital heart disease." eng. In: *J Med Genet* 45.11 (Nov. 2008), pp. 704–709. DOI: 10.1136/jmg.2008.058776. URL: http://dx.doi.org/10.1136/jmg.2008.058776.

[154] Evangelia Karampetsou, Deborah Morrogh, and Lyn Chitty. "Microarray Technology for the Diagnosis of Fetal Chromosomal Aberrations: Which Platform Should We Use?" eng. In: *J Clin Med* 3.2 (2014), pp. 663–678. DOI: 10.3390/jcm3020663. URL: http://dx.doi.org/10.3390/jcm3020663.

[155] Chao Xie and Martti T. Tammi. "CNV-seq, a new method to detect copy number variation using high-throughput sequencing." eng. In: *BMC Bioinformatics* 10 (2009), p. 80. DOI: 10.1186/1471-2105-10-80. URL: http://dx.doi.org/10.1186/1471-2105-10-80.

[156] Seungtai Yoon et al. "Sensitive and accurate detection of copy number variants using read depth of coverage." eng. In: *Genome Res* 19.9 (Sept. 2009), pp. 1586–1592. DOI: 10.1101/gr.092981.109. URL: http://dx.doi.org/10.1101/gr.092981.109.

[157] Can Alkan, Bradley P. Coe, and Evan E. Eichler. "Genome structural variation discovery and genotyping." eng. In: *Nat Rev Genet* 12.5 (May 2011), pp. 363–376. DOI: 10.1038/nrg2958. URL: http://dx.doi.org/10.1038/nrg2958.

[158] Yuhao Shi and Jacek Majewski. "FishingCNV: a graphical software package for detecting rare copy number variations in exome-sequencing data." eng. In: *Bioinformatics* 29.11 (June 2013), pp. 1461–1462. DOI: 10.1093/bioinformatics/btt151.

[159] Jason Li et al. "CONTRA: copy number analysis for targeted resequencing." eng. In: *Bioinformatics* 28.10 (May 2012), pp. 1307–1313. DOI: 10.1093/bioinformatics/bts146. URL: http://dx.doi.org/10.1093/bioinformatics/bts146.

[160] Jarupon Fah Sathirapongsasuti et al. "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV." eng. In: *Bioinformatics* 27.19 (Oct. 2011), pp. 2648–2654. DOI: 10.1093/bioinformatics/btr462. URL: http://dx.doi.org/10.1093/bioinformatics/btr462.

[161] Vincent Plagnol et al. "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling." eng. In: *Bioinformatics* 28.21 (Nov. 2012), pp. 2747–2754. URL: http://dx.doi.org/10.1093/bioinformatics/bts526.

[162] Menachem Fromer et al. "Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth." eng. In: *Am J Hum Genet* 91.4 (Oct. 2012), pp. 597–607. DOI: 10.1016/j.ajhg.2012.08.005. URL: http://dx.doi.org/10.1016/j.ajhg.2012.08.005.

[163] Kaushalya C. Amarasinghe, Jason Li, and Saman K. Halgamuge. "CoNVEX: copy number variation estimation in exome sequencing data using HMM." eng. In: *BMC Bioinformatics* 14 Suppl 2 (2013), S2. DOI: 10.1186/1471-2105-14-S2-S2. URL: http://dx.doi.org/10.1186/1471-2105-14-S2-S2.

[164] Christopher A. Miller et al. "ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads." eng. In: *PLoS One* 6.1 (2011), e16327. DOI: 10.1371/journal.pone.0016327. URL: http://dx.doi.org/10.1371/journal.pone.0016327.

[165] Alexej Abyzov et al. "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing." eng. In: *Genome Res* 21.6 (June 2011), pp. 974–984. DOI: 10.1101/gr.114876.110. URL: http://dx.doi.org/10.1101/gr.114876.110.

[166] Arief Gusnanto et al. "Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data." eng. In: *Bioinformatics* 28.1 (Jan. 2012), pp. 40–47. DOI: 10.1093/bioinformatics/btr593. URL: http://dx.doi.org/10.1093/bioinformatics/btr593.

[167] Niklas Krumm et al. "Copy number variation detection and genotyping from exome sequence data." eng. In: *Genome Res* 22.8 (Aug. 2012), pp. 1525–1532. DOI: 10.1101/gr.138115.112. URL: http://dx.doi.org/10.1101/gr.138115.112.

[168] Yuval Benjamini and Terence P. Speed. "Summarizing and correcting the GC content bias in high-throughput sequencing." eng. In: *Nucleic Acids Res* 40.10 (May 2012), e72. DOI: 10.1093/nar/gks001. URL: http://dx.doi.org/10.1093/nar/gks001.

[169] Valentina Boeva et al. “Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization.” eng. In: *Bioinformatics* 27.2 (Jan. 2011), pp. 268–269. DOI: 10.1093/bioinformatics/btq635. URL: http://dx.doi.org/10.1093/bioinformatics/btq635.

[170] *micro-read Copy Number Variant Regions*. http://mrcanavar.sourceforge.net/.

[171] Michael I Love et al. “Modeling read counts for CNV detection in exome sequencing data”. In: *Statistical applications in genetics and molecular biology* 10.1 (2011).

[172] Peter V. Kharchenko, Michael Y. Tolstorukov, and Peter J. Park. “Design and analysis of ChIP-seq experiments for DNA-binding proteins.” eng. In: *Nat Biotechnol* 26.12 (Dec. 2008), pp. 1351–1359. DOI: 10.1038/nbt.1508. URL: http://dx.doi.org/10.1038/nbt.1508.

[173] Joel Rozowsky et al. “PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.” eng. In: *Nat Biotechnol* 27.1 (Jan. 2009), pp. 66–75. DOI: 10.1038/nbt.1518. URL: http://dx.doi.org/10.1038/nbt.1518.

[174] Hongkai Ji et al. “An integrated software system for analyzing ChIP-chip and ChIP-seq data.” eng. In: *Nat Biotechnol* 26.11 (Nov. 2008), pp. 1293–1300. DOI: 10.1038/nbt.1505. URL: http://dx.doi.org/10.1038/nbt.1505.

[175] Yong Zhang et al. “Model-based analysis of ChIP-Seq (MACS).” eng. In: *Genome Biol* 9.9 (2008), R137. DOI: 10.1186/gb-2008-9-9-r137. URL: http://dx.doi.org/10.1186/gb-2008-9-9-r137.

[176] Mahmoud M. Ibrahim, Scott A. Lacadie, and Uwe Ohler. “JAMM: a peak finder for joint analysis of NGS replicates.” eng. In: *Bioinformatics* 31.1 (Jan. 2015), pp. 48–55. DOI: 10.1093/bioinformatics/btu568. URL: http://dx.doi.org/10.1093/bioinformatics/btu568.

[177] Jianxing Feng, Tao Liu, and Yong Zhang. “Using MACS to identify peaks from ChIP-Seq data.” eng. In: *Curr Protoc Bioinformatics* Chapter 2 (June 2011), Unit 2.14. DOI: 10.1002/0471250953.bi0214s34. URL: http://dx.doi.org/10.1002/0471250953.bi0214s34.

[178] Shirley Pepke, Barbara Wold, and Ali Mortazavi. “Computation for ChIP-seq and RNA-seq studies.” eng. In: *Nat Methods* 6.11 Suppl (Nov. 2009), S22–S32. DOI: 10.1038/nmeth.1371. URL: http://dx.doi.org/10.1038/nmeth.1371.

[179] David Sims et al. “Sequencing depth and coverage: key considerations in genomic analyses.” eng. In: *Nat Rev Genet* 15.2 (Feb. 2014), pp. 121–132. DOI: 10.1038/nrg3642. URL: http://dx.doi.org/10.1038/nrg3642.

[180] *MACS (2.1.0)*. https://github.com/taoliu/MACS.

[181] Timothy Bailey et al. “Practical guidelines for the comprehensive analysis of ChIP-seq data.” eng. In: *PLoS Comput Biol* 9.11 (2013), e1003326. DOI: 10.1371/journal.pcbi.1003326. URL: http://dx.doi.org/10.1371/journal.pcbi.1003326.

[182] V. Matys et al. “TRANSFAC: transcriptional regulation, from patterns to profiles.” eng. In: *Nucleic Acids Res* 31.1 (Jan. 2003), pp. 374–378.

[183] Morgane Thomas-Chollier et al. “A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs.” eng. In: *Nat Protoc* 7.8 (Aug. 2012), pp. 1551–1568. DOI: 10.1038/nprot.2012.088. URL: http://dx.doi.org/10.1038/nprot.2012.088.

[184] Morgane Thomas-Chollier et al. “RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets.” eng. In: *Nucleic Acids Res* 40.4 (Feb. 2012), e31. DOI: 10.1093/nar/gkr1104. URL: http://dx.doi.org/10.1093/nar/gkr1104.

[185] A. E. Kel et al. “MATCH: A tool for searching transcription factor binding sites in DNA sequences.” eng. In: *Nucleic Acids Res* 31.13 (July 2003), pp. 3576–3579.

[186] T. D. Schneider and R. M. Stephens. “Sequence logos: a new way to display consensus sequences.” eng. In: *Nucleic Acids Res* 18.20 (Oct. 1990), pp. 6097–6100.

[187] Anthony Mathelier et al. “JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.” eng. In: *Nucleic Acids Res* 42.Database issue (Jan. 2014), pp. D142–D147. DOI: 10.1093/nar/gkt997. URL: http://dx.doi.org/10.1093/nar/gkt997.

[188] Morgane Thomas-Chollier et al. “RSAT 2011: regulatory sequence analysis tools.” eng. In: *Nucleic Acids Res* 39.Web Server issue (July 2011), W86–W91. DOI: 10.1093/nar/gkr377. URL: http://dx.doi.org/10.1093/nar/gkr377.

[189] Alejandra Medina-Rivera et al. “RSAT 2015: Regulatory Sequence Analysis Tools.” eng. In: *Nucleic Acids Res* 43.W1 (July 2015), W50–W56. DOI: 10.1093/nar/gkv362. URL: http://dx.doi.org/10.1093/nar/gkv362.

[190] M. Ashburner et al. “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.” eng. In: *Nat Genet* 25.1 (May 2000), pp. 25–29. DOI: 10.1038/75556. URL: http://dx.doi.org/10.1038/75556.

[191] Judith A. Blake. “Ten quick tips for using the gene ontology.” eng. In: *PLoS Comput Biol* 9.11 (2013), e1003343. DOI: 10.1371/journal.pcbi.1003343. URL: http://dx.doi.org/10.1371/journal.pcbi.1003343.

[192] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.” eng. In: *Nat Protoc* 4.1 (2009), pp. 44–57. DOI: 10.1038/nprot.2008.211. URL: http://dx.doi.org/10.1038/nprot.2008.211.

[193] Judith A. Blake et al. “The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse.” eng. In: *Nucleic Acids Res* 42.Database issue (Jan. 2014), pp. D810–D817. DOI: 10.1093/nar/gkt1225. URL: http://dx.doi.org/10.1093/nar/gkt1225.

[194] Aleksandra Pekowska et al. "A unique H3K4me2 profile marks tissue-specific gene regulation." eng. In: *Genome Res* 20.11 (Nov. 2010), pp. 1493–1502. DOI: 10.1101/gr.109389.110. URL: http://dx.doi.org/10.1101/gr.109389.110.

[195] Jie Zhang, Jeffrey Parvin, and Kun Huang. "Redistribution of H3K4me2 on neural tissue specific genes during mouse brain development." eng. In: *BMC Genomics* 13 Suppl 8 (2012), S5. DOI: 10.1186/1471-2164-13-S8-S5. URL: http://dx.doi.org/10.1186/1471-2164-13-S8-S5.

[196] Evgenya Y. Popova et al. "Stage and gene specific signatures defined by histones H3K4me2 and H3K27me3 accompany mammalian retina maturation in vivo." eng. In: *PLoS One* 7.10 (2012), e46867. DOI: 10.1371/journal.pone.0046867. URL: http://dx.doi.org/10.1371/journal.pone.0046867.

[197] Marie-Agnes Dillies and Andrea Rau. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis." eng. In: *Brief Bioinform* 14.6 (Nov. 2013), pp. 671–683.

[198] Cole Trapnell et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." eng. In: *Nat Biotechnol* 28.5 (May 2010), pp. 511–515. DOI: 10.1038/nbt.1621. URL: http://dx.doi.org/10.1038/nbt.1621.

[199] Cole Trapnell et al. "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." eng. In: *Nat Protoc* 7.3 (Mar. 2012), pp. 562–578. DOI: 10.1038/nprot.2012.016. URL: http://dx.doi.org/10.1038/nprot.2012.016.

[200] Daniel Hebenstreit et al. "RNA sequencing reveals two major classes of gene expression levels in metazoan cells." eng. In: *Mol Syst Biol* 7 (2011), p. 497. DOI: 10.1038/msb.2011.28. URL: http://dx.doi.org/10.1038/msb.2011.28.

[201] Cole Trapnell et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." eng. In: *Nat Biotechnol* 31.1 (Jan. 2013), pp. 46–53. DOI: 10.1038/nbt.2450. URL: http://dx.doi.org/10.1038/nbt.2450.

[202] Alicia Oshlack, Mark D. Robinson, and Matthew D. Young. "From RNA-seq reads to differential expression results." eng. In: *Genome Biol* 11.12 (2010), p. 220. DOI: 10.1186/gb-2010-11-12-220. URL: http://dx.doi.org/10.1186/gb-2010-11-12-220.

[203] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data." eng. In: *Genome Biol* 11.10 (2010), R106. DOI: 10.1186/gb-2010-11-10-r106. URL: http://dx.doi.org/10.1186/gb-2010-11-10-r106.

[204] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." eng. In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140. DOI: 10.1093/bioinformatics/btp616. URL: http://dx.doi.org/10.1093/bioinformatics/btp616.

[205] David Hiller et al. "Identifiability of isoform deconvolution from junction arrays and RNA-Seq." eng. In: *Bioinformatics* 25.23 (Dec. 2009), pp. 3056–3059. DOI: 10.1093/bioinformatics/btp544. URL: http://dx.doi.org/10.1093/bioinformatics/btp544.

[206] Thomas J. Hardcastle and Krystyna A. Kelly. "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data." eng. In: *BMC Bioinformatics* 11 (2010), p. 422. DOI: 10.1186/1471-2105-11-422. URL: http://dx.doi.org/10.1186/1471-2105-11-422.

[207] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." eng. In: *Genome Biol* 15.12 (2014), p. 550. DOI: 10.1186/s13059-014-0550-8. URL: http://dx.doi.org/10.1186/s13059-014-0550-8.

[208] Carlo E Bonferroni. *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber, 1936.

[209] R. Bender and S. Lange. "Multiple test procedures other than Bonferroni's deserve wider use." eng. In: *BMJ* 318.7183 (Feb. 1999), pp. 600–601.

[210] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1995), pp. 289–300.

[211] Eric T. Wang et al. "Alternative isoform regulation in human tissue transcriptomes." eng. In: *Nature* 456.7221 (Nov. 2008), pp. 470–476. DOI: 10.1038/nature07509. URL: http://dx.doi.org/10.1038/nature07509.

[212] Qun Pan et al. "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." eng. In: *Nat Genet* 40.12 (Dec. 2008), pp. 1413–1415. DOI: 10.1038/ng.259. URL: http://dx.doi.org/10.1038/ng.259.

[213] Michael G. Poulos et al. "Developments in RNA splicing and disease." eng. In: *Cold Spring Harb Perspect Biol* 3.1 (Jan. 2011), a000778. DOI: 10.1101/cshperspect.a000778. URL: http://dx.doi.org/10.1101/cshperspect.a000778.

[214] Jamal Tazi, Nadia Bakkour, and Stefan Stamm. "Alternative splicing and disease." eng. In: *Biochim Biophys Acta* 1792.1 (Jan. 2009), pp. 14–26. DOI: 10.1016/j.bbadis.2008.09.017. URL: http://dx.doi.org/10.1016/j.bbadis.2008.09.017.

[215] Andrew G L. Douglas and Matthew J A. Wood. "RNA splicing: disease and therapy." eng. In: *Brief Funct Genomics* 10.3 (May 2011), pp. 151–164. DOI: 10.1093/bfgp/elr020. URL: http://dx.doi.org/10.1093/bfgp/elr020.

[216] Xuexia Zhou et al. "Transcriptome analysis of alternative splicing events regulated by SRSF10 reveals position-dependent splicing modulation." eng. In: *Nucleic Acids Res* 42.6 (Apr. 2014), pp. 4019–4030. DOI: 10.1093/nar/gkt1387. URL: http://dx.doi.org/10.1093/nar/gkt1387.

[217] Simon Anders, Alejandro Reyes, and Wolfgang Huber. "Detecting differential usage of exons from RNA-seq data." eng. In: *Genome Res* 22.10 (Oct. 2012), pp. 2008–2017. DOI: 10.1101/gr.133744.111. URL: http://dx.doi.org/10.1101/gr.133744.111.

[218] Yarden Katz et al. "Analysis and design of RNA sequencing experiments for identifying isoform regulation." eng. In: *Nat Methods* 7.12 (Dec. 2010), pp. 1009–1015. DOI: 10.1038/nmeth.1528. URL: http://dx.doi.org/10.1038/nmeth.1528.

[219] Dmitri D. Pervouchine, David G. Knowles, and Roderic Guigo. "Intron-centric estimation of alternative splicing from RNA-seq data." eng. In: *Bioinformatics* 29.2 (Jan. 2013), pp. 273–274. DOI: 10.1093/bioinformatics/bts678. URL: http://dx.doi.org/10.1093/bioinformatics/bts678.

[220] Boyko Kakaradov et al. "Challenges in estimating percent inclusion of alternatively spliced junctions from RNA-seq data." eng. In: *BMC Bioinformatics* 13 Suppl 6 (2012), S11. DOI: 10.1186/1471-2105-13-S6-S11. URL: http://dx.doi.org/10.1186/1471-2105-13-S6-S11.

[221] Wei Guo et al. "RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing." eng. In: *Nat Med* 18.5 (May 2012), pp. 766–773. DOI: 10.1038/nm.2693. URL: http://dx.doi.org/10.1038/nm.2693.

[222] Stefanie Hammer et al. "Characterization of TBX20 in human hearts and its regulation by TFAP2." eng. In: *J Cell Biochem* 104.3 (June 2008), pp. 1022–1033. DOI: 10.1002/jcb.21686. URL: http://dx.doi.org/10.1002/jcb.21686.

[223] Martje Toenjes et al. "Prediction of cardiac transcription networks based on molecular data and complex clinical phenotypes." eng. In: *Mol Biosyst* 4.6 (June 2008), pp. 589–598. DOI: 10.1039/b800207j. URL: http://dx.doi.org/10.1039/b800207j.

[224] International HapMap. "Integrating common and rare genetic variation in diverse human populations." eng. In: *Nature* 467.7311 (Sept. 2010), pp. 52–58. DOI: 10.1038/nature09298. URL: http://dx.doi.org/10.1038/nature09298.

[225] D. Yaffe and O. Saxel. "Serial passaging and differentiation of myogenic cells isolated from dystrophic mouse muscle." eng. In: *Nature* 270.5639 (1977), pp. 725–727.

[226] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. "Structural variation in the human genome." eng. In: *Nat Rev Genet* 7.2 (Feb. 2006), pp. 85–97. DOI: 10.1038/nrg1767. URL: http://dx.doi.org/10.1038/nrg1767.

[227] Donald F. Conrad et al. "Origins and functional impact of copy number variation in the human genome." eng. In: *Nature* 464.7289 (Apr. 2010), pp. 704–712. DOI: 10.1038/nature08516. URL: http://dx.doi.org/10.1038/nature08516.

[228] W. J. Dixon. "Analysis of Extreme Values". In: *Ann. Math. Statist.* 21.4 (Dec. 1950), pp. 488–506. DOI: 10.1214/aoms/1177729747. URL: http://dx.doi.org/10.1214/aoms/1177729747.

[229] David B Rorabacher. "Statistical treatment for rejection of deviant values: critical values of Dixon's" Q" parameter and related subrange ratios at the 95% confidence level". In: *Analytical Chemistry* 63.2 (1991), pp. 139–146.

[230] Lawrence R. Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *PROCEEDINGS OF THE IEEE*. 1989, pp. 257–286.

[231] Leonard E Baum et al. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *The annals of mathematical statistics* (1970), pp. 164–171.

[232] Andrew J Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *Information Theory, IEEE Transactions on* 13.2 (1967), pp. 260–269.

[233] Maria V. Neguembor and Davide Gabellini. "In junk we trust: repetitive DNA, epigenetics and facioscapulohumeral muscular dystrophy." eng. In: *Epigenomics* 2.2 (Apr. 2010), pp. 271–287. DOI: 10.2217/epi.10.8. URL: http://dx.doi.org/10.2217/epi.10.8.

[234] Silvia Consalvi et al. "Histone deacetylase inhibitors in the treatment of muscular dystrophies: epigenetic drugs for genetic diseases." eng. In: *Mol Med* 17.5-6 (2011), pp. 457–465. DOI: 10.2119/molmed.2011.00049. URL: http://dx.doi.org/10.2119/molmed.2011.00049.

[235] Margaret Buckingham and Peter W J. Rigby. "Gene regulatory networks and transcriptional mechanisms that control myogenesis." eng. In: *Dev Cell* 28.3 (Feb. 2014), pp. 225–238. DOI: 10.1016/j.devcel.2013.12.020. URL: http://dx.doi.org/10.1016/j.devcel.2013.12.020.

[236] Frédéric Relaix et al. "Six homeoproteins directly activate Myod expression in the gene regulatory networks that control early myogenesis." eng. In: *PLoS Genet* 9.4 (Apr. 2013), e1003425. DOI: 10.1371/journal.pgen.1003425. URL: http://dx.doi.org/10.1371/journal.pgen.1003425.

[237] R. E. Breitbart et al. "A fourth human MEF2 transcription factor, hMEF2D, is an early marker of the myogenic lineage." eng. In: *Development* 118.4 (Aug. 1993), pp. 1095–1106.

[238] Ning Liu et al. "Requirement of MEF2A, C, and D for skeletal muscle regeneration." eng. In: *Proc Natl Acad Sci U S A* 111.11 (Mar. 2014), pp. 4109–4114. DOI: 10.1073/pnas.1401732111. URL: http://dx.doi.org/10.1073/pnas.1401732111.

[239] Charis L. Himeda et al. "KLF3 regulates muscle-specific gene expression and synergizes with serum response factor on KLF binding sites." eng. In: *Mol Cell Biol* 30.14 (July 2010), pp. 3430–3443. DOI: 10.1128/MCB.00302-10. URL: http://dx.doi.org/10.1128/MCB.00302-10.

[240] Ravi Chandran et al. "Biomechanical signals upregulate myogenic gene induction in the presence or absence of inflammation." eng. In: *Am J Physiol Cell Physiol* 293.1 (July 2007), pp. C267–C276. DOI: 10.1152/ajpcell.00594.2006. URL: http://dx.doi.org/10.1152/ajpcell.00594.2006.

[241] D. C. Rockey, C. N. Housset, and S. L. Friedman. "Activation-dependent contractility of rat hepatic lipocytes in culture and in vivo." eng. In: *J Clin Invest* 92.4 (Oct. 1993), pp. 1795–1804. DOI: 10.1172/JCI116769. URL: http://dx.doi.org/10.1172/JCI116769.

[242] Saravanan Rajan et al. "Analysis of early C2C12 myogenesis identifies stably and differentially expressed transcriptional regulators whose knock-down inhibits myoblast differentiation." eng. In: *Physiol Genomics* 44.2 (Feb. 2012), pp. 183–197. DOI: 10.1152/physiolgenomics.00093.2011. URL: http://dx.doi.org/10.1152/physiolgenomics.00093.2011.

[243] Matthew S. Hestand et al. "Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies." eng. In: *Nucleic Acids Res* 38.16 (Sept. 2010), e165. DOI: 10.1093/nar/gkq602. URL: http://dx.doi.org/10.1093/nar/gkq602.

[244] H. Weintraub et al. "Muscle-specific transcriptional activation by MyoD." eng. In: *Genes Dev* 5.8 (Aug. 1991), pp. 1377–1386.

[245] E. N. Olson. "Regulation of muscle transcription by the MyoD family. The heart of the matter." eng. In: *Circ Res* 72.1 (Jan. 1993), pp. 1–6.

[246] Jin Rong Ow et al. "Patz1 regulates embryonic stem cell identity." eng. In: *Stem Cells Dev* 23.10 (May 2014), pp. 1062–1073. DOI: 10.1089/scd.2013.0430. URL: http://dx.doi.org/10.1089/scd.2013.0430.

[247] Keith Orford et al. "Differential H3K4 methylation identifies developmentally poised hematopoietic genes." eng. In: *Dev Cell* 14.5 (May 2008), pp. 798–809. DOI: 10.1016/j.devcel.2008.04.002. URL: http://dx.doi.org/10.1016/j.devcel.2008.04.002.

[248] Reini F. Luco et al. "Epigenetics in alternative pre-mRNA splicing." eng. In: *Cell* 144.1 (Jan. 2011), pp. 16–26. DOI: 10.1016/j.cell.2010.11.056. URL: http://dx.doi.org/10.1016/j.cell.2010.11.056.

[249] *AltAnalyze*. https://code.google.com/p/altanalyze/wiki/.

[250] L. Fananapazir et al. “Missense mutations in the beta-myosin heavy-chain gene cause central core disease in hypertrophic cardiomyopathy.” eng. In: *Proc Natl Acad Sci U S A* 90.9 (May 1993), pp. 3993–3997.

[251] Eric Villard et al. “Mutation screening in dilated cardiomyopathy: prominent role of the beta myosin heavy chain gene.” eng. In: *Eur Heart J* 26.8 (Apr. 2005), pp. 794–803. DOI: 10.1093/eurheartj/ehi193. URL: http://dx.doi.org/10.1093/eurheartj/ehi193.

[252] Lei Bu et al. “Human ISL1 heart progenitors generate diverse multipotent cardiovascular cell lineages.” eng. In: *Nature* 460.7251 (July 2009), pp. 113–117. DOI: 10.1038/nature08191. URL: http://dx.doi.org/10.1038/nature08191.

[253] Kristen N. Stevens et al. “Common variation in ISL1 confers genetic susceptibility for human congenital heart disease.” eng. In: *PLoS One* 5.5 (2010), e10855. DOI: 10.1371/journal.pone.0010855. URL: http://dx.doi.org/10.1371/journal.pone.0010855.

[254] Jione Kang et al. “Isl1 is a direct transcriptional target of Forkhead transcription factors in second-heart-field-derived mesoderm.” eng. In: *Dev Biol* 334.2 (Oct. 2009), pp. 513–522. DOI: 10.1016/j.ydbio.2009.06.041. URL: http://dx.doi.org/10.1016/j.ydbio.2009.06.041.

[255] Mohamed Nemir and Thierry Pedrazzini. “Functional role of Notch signaling in the developing and postnatal heart.” eng. In: *J Mol Cell Cardiol* 45.4 (Oct. 2008), pp. 495–504. DOI: 10.1016/j.yjmcc.2008.02.273. URL: http://dx.doi.org/10.1016/j.yjmcc.2008.02.273.

[256] Vidu Garg et al. “Mutations in NOTCH1 cause aortic valve disease.” eng. In: *Nature* 437.7056 (Sept. 2005), pp. 270–274. DOI: 10.1038/nature03940. URL: http://dx.doi.org/10.1038/nature03940.

[257] Salah A. Mohamed et al. “Novel missense mutations (p.T596M and p.P1797H) in NOTCH1 in patients with bicuspid aortic valve.” eng. In: *Biochem Biophys Res Commun* 345.4 (July 2006), pp. 1460–1465. DOI: 10.1016/j.bbrc.2006.05.046. URL: http://dx.doi.org/10.1016/j.bbrc.2006.05.046.

[258] Kazuo Momma. “Cardiovascular anomalies associated with chr 22q11.2 deletion syndrome.” eng. In: *Am J Cardiol* 105.11 (June 2010), pp. 1617–1624. DOI: 10.1016/j.amjcard.2010.01.333. URL: http://dx.doi.org/10.1016/j.amjcard.2010.01.333.

[259] Gary Humphreys. “Coming together to combat rare diseases.” eng. In: *Bull World Health Organ* 90.6 (June 2012), pp. 406–407. DOI: 10.2471/BLT.12.020612. URL: http://dx.doi.org/10.2471/BLT.12.020612.

[260] Vahab D. Soleimani et al. “Snail regulates MyoD binding-site occupancy to direct enhancer switching and differentiation-specific transcription in myogenesis.” eng. In: *Mol Cell* 47.3 (Aug. 2012), pp. 457–468. DOI: 10.1016/j.molcel.2012.05.046. URL: http://dx.doi.org/10.1016/j.molcel.2012.05.046.

[261] L. A. Groom et al. "Differential regulation of the MAP, SAP and RK/p38 kinases by Pyst1, a novel cytosolic dual-specificity phosphatase." eng. In: *EMBO J* 15.14 (July 1996), pp. 3621–3632.

[262] J. D. Molkentin et al. "Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins." eng. In: *Cell* 83.7 (Dec. 1995), pp. 1125–1136.

[263] Stephanie Wales et al. "Global MEF2 target gene analysis in cardiac and skeletal muscle reveals novel regulation of DUSP6 by p38MAPK-MEF2 signaling." eng. In: *Nucleic Acids Res* 42.18 (Oct. 2014), pp. 11349–11362. DOI: 10.1093/nar/gku813. URL: http://dx.doi.org/10.1093/nar/gku813.

[264] Jung-Chun Lin and Woan-Yuh Tarn. "Exon selection in alpha-tropomyosin mRNA is regulated by the antagonistic action of RBM4 and PTB." eng. In: *Mol Cell Biol* 25.22 (Nov. 2005), pp. 10111–10121. DOI: 10.1128/MCB.25.22.10111-10121.2005. URL: http://dx.doi.org/10.1128/MCB.25.22.10111-10121.2005.

[265] T. Valentino et al. "PATZ1 interacts with p53 and regulates expression of p53-target genes enhancing apoptosis or cell survival based on the cellular context." eng. In: *Cell Death Dis* 4 (2013), e963. DOI: 10.1038/cddis.2013.500. URL: http://dx.doi.org/10.1038/cddis.2013.500.

[266] Teresa Valentino et al. "Embryonic defects and growth alteration in mice with homozygous disruption of the Patz1 gene." eng. In: *J Cell Physiol* 228.3 (Mar. 2013), pp. 646–653. DOI: 10.1002/jcp.24174. URL: http://dx.doi.org/10.1002/jcp.24174.

[267] Jennifer C J. Chen et al. "MyoD-cre transgenic mice: a model for conditional mutagenesis and lineage tracing of skeletal muscle." eng. In: *Genesis* 41.3 (Mar. 2005), pp. 116–121. DOI: 10.1002/gene.20104. URL: http://dx.doi.org/10.1002/gene.20104.

[268] Hadas Keren, Galit Lev-Maor, and Gil Ast. "Alternative splicing and evolution: diversification, exon definition and function." eng. In: *Nat Rev Genet* 11.5 (May 2010), pp. 345–355. DOI: 10.1038/nrg2776. URL: http://dx.doi.org/10.1038/nrg2776.

[269] Alexander V. Alekseyenko, Namshin Kim, and Christopher J. Lee. "Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes." eng. In: *RNA* 13.5 (May 2007), pp. 661–670. DOI: 10.1261/rna.325107. URL: http://dx.doi.org/10.1261/rna.325107.

[270] C. W. Sugnet et al. "Transcriptome and genome conservation of alternative splicing events in humans and mice." eng. In: *Pac Symp Biocomput* (2004), pp. 66–77.

[271] Eddo Kim, Amir Goren, and Gil Ast. "Alternative splicing: current perspectives." eng. In: *Bioessays* 30.1 (Jan. 2008), pp. 38–47. DOI: 10.1002/bies.20692. URL: http://dx.doi.org/10.1002/bies.20692.

[272] Noboru Jo Sakabe and Sandro José de Souza. "Sequence features responsible for intron retention in human." eng. In: *BMC Genomics* 8 (2007), p. 59. DOI: 10.1186/1471-2164-8-59. URL: http://dx.doi.org/10.1186/1471-2164-8-59.

[273] Gil Ast. "How did alternative splicing evolve?" eng. In: *Nat Rev Genet* 5.10 (Oct. 2004), pp. 773–782. DOI: 10.1038/nrg1451. URL: http://dx.doi.org/10.1038/nrg1451.

[274] Douglas L. Black. "Mechanisms of alternative pre-messenger RNA splicing." eng. In: *Annu Rev Biochem* 72 (2003), pp. 291–336. DOI: 10.1146/annurev.biochem.72.121801.161720. URL: http://dx.doi.org/10.1146/annurev.biochem.72.121801.161720.

[275] Yuwen Liu, Jie Zhou, and Kevin P. White. "RNA-seq differential expression studies: more sequence or more replication?" eng. In: *Bioinformatics* 30.3 (Feb. 2014), pp. 301–304. DOI: 10.1093/bioinformatics/btt688. URL: http://dx.doi.org/10.1093/bioinformatics/btt688.

[276] Shihao Shen et al. "rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data." eng. In: *Proc Natl Acad Sci U S A* 111.51 (Dec. 2014), E5593–E5601. DOI: 10.1073/pnas.1419161111. URL: http://dx.doi.org/10.1073/pnas.1419161111.

[277] Ho Sung Rhee and B Franklin Pugh. "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution." eng. In: *Cell* 147.6 (Dec. 2011), pp. 1408–1419. DOI: 10.1016/j.cell.2011.11.013. URL: http://dx.doi.org/10.1016/j.cell.2011.11.013.

[278] Ho Sung Rhee and B Franklin Pugh. "ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy." eng. In: *Curr Protoc Mol Biol* Chapter 21 (Oct. 2012), Unit 21.24. DOI: 10.1002/0471142727.mb2124s100. URL: http://dx.doi.org/10.1002/0471142727.mb2124s100.

[279] Ho Sung Rhee and B Franklin Pugh. "Genome-wide structure and organization of eukaryotic pre-initiation complexes." eng. In: *Nature* 483.7389 (Mar. 2012), pp. 295–301. DOI: 10.1038/nature10799. URL: http://dx.doi.org/10.1038/nature10799.

[280] Yuchun Guo, Shaun Mahony, and David K. Gifford. "High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints." eng. In: *PLoS Comput Biol* 8.8 (2012), e1002638. DOI: 10.1371/journal.pcbi.1002638. URL: http://dx.doi.org/10.1371/journal.pcbi.1002638.

[281] Stephan R. Starick et al. "ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors." eng. In: *Genome Res* 25.6 (June 2015), pp. 825–835. DOI: 10.1101/gr.185157.114. URL: http://dx.doi.org/10.1101/gr.185157.114.

[282] Aurelien A. Serandour et al. "Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties." eng. In: *Genome Biol* 14.12 (2013), R147. DOI: 10.1186/gb-2013-14-12-r147. URL: http://dx.doi.org/10.1186/gb-2013-14-12-r147.

[283] Ho Sung Rhee et al. "Subnucleosomal structures and nucleosome asymmetry across a genome." eng. In: *Cell* 159.6 (Dec. 2014), pp. 1377–1388. DOI: 10.1016/j.cell.2014.10.054. URL: http://dx.doi.org/10.1016/j.cell.2014.10.054.

[284] Günter P. Wagner, Koryu Kin, and Vincent J. Lynch. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples." eng. In: *Theory Biosci* 131.4 (Dec. 2012), pp. 281–285. DOI: 10.1007/s12064-012-0162-3. URL: http://dx.doi.org/10.1007/s12064-012-0162-3.

[285] *What the FPKM? A review of RNA-Seq expression units.* URL: https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/.

[286] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. "Single-cell sequencing-based technologies will revolutionize whole-organism science." eng. In: *Nat Rev Genet* 14.9 (Sept. 2013), pp. 618–630. DOI: 10.1038/nrg3542. URL: http://dx.doi.org/10.1038/nrg3542.

[287] Angela R. Wu et al. "Quantitative assessment of single-cell RNA-sequencing methods." eng. In: *Nat Methods* 11.1 (Jan. 2014), pp. 41–46. DOI: 10.1038/nmeth.2694. URL: http://dx.doi.org/10.1038/nmeth.2694.

[288] Evan Z. Macosko et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets." eng. In: *Cell* 161.5 (May 2015), pp. 1202–1214. DOI: 10.1016/j.cell.2015.05.002. URL: http://dx.doi.org/10.1016/j.cell.2015.05.002.

[289] *New Challenges in the Fast Changing Landscape of Bioinformatics.* http://www.biotech-now.org/health/2013/04/new-challenges-in-the-fast-changing-landscape-of-bioinformatics/.

[290] Vivien Marx. "Biology: The big challenges of big data". In: *Nature* 498.7453 (2013), pp. 255–260.

Zusammenfassung

Mit der Technologie der Hochdurchsatzsequenzierung (HTS) hat sich die Forschung in den Lebenswissenschaften in den letzten Jahren stark beschleunigt. Aufgrund der niedrigen Kosten und der hohen Effizienz wird diese Technologie heutzutage häufig zur Beantwortung verschiedenster biologischer Fragestellungen verwendet. Im Allgemeinen wird bei der HTS die Sequenz von Millionen von DNA-Fragmenten parallel bestimmt wobei diese Fragmente wiederum unter Verwendung unterschiedlicher Sequenzierungsverfahren erzeugt werden können. Mit den schnellen Fortschritten in dieser Technologie scheinen deren Anwendungen fast unbegrenzt. Beispielsweise ist es nun möglich, ein ganzes Genom in weniger als einem Tag komplett zu sequenzieren. Neben der Sequenzierung ganzer Genome gibt es noch verschiedene andere Anwendungen, wie die gezielte Resequenzierung von DNA, die Quantifizierung von Genexpressionsprofilen (RNA-seq) und die genomweite Identifikation von Protein-DNA-Wechselwirkungen, wie beispielsweise Transkriptionsfaktor-Bindungsstellen oder Chromatin-Histon-Modifikationen (ChIP-seq). Die Analyse der durch HTS erzeugten massiven Datenmengen erfordert allerdings auch ausgefeilte bioinformatorische Methoden. In dieser Dissertation präsentiere ich computerbasierte Ansätze zur Analyse von gezielten DNA-Resequenzierungsdaten sowie von RNA-seq- und ChIP-seq-Daten, um biologische Fragen hinsichtlich Herzerkrankungen und Skelettmuskelentwicklung zu beantworten.

Im ersten Teil der Arbeit wurde eine neue Methode zur Identifizierung von individuellen, krankheitsrelevanten Kopienzahlvariationen (engl. *copy number variations*, kurz CNVs) unter Verwendung von gezielten Resequenzierungs- oder Exomdaten mit kleinem Probenumfang entwickelt. Das Auffinden von CNVs in gezielten Resequenzierungsdaten ist aufgrund ungleichmäßiger Lesetiefen zwischen den erfassten Regionen schwierig. Darüber hinaus wurde eine Methode gebraucht, um individuelle CNVs von einer kleinen Patientenkohorte zu erfassen, ohne entsprechende Kontrollen zu verwenden. Wir haben daher eine solche Methode entwickelt, die wir basierend auf den verfügbaren Daten von acht Proben aus dem HapMap Projekt sowie auf einer kleinen Anzahl von Patienten mit Fallot'scher Tetralogie (TOF) untersucht haben. Zusätzlich zu unserer Methode haben wir bei der Evaluierung die beiden frei zugänglichen Programme ExomeDepth und CoNIFER verwendet. Dabei zeigte sich, dass ExomeDepth im Vergleich zu CoNIFER und unserer Methode mehr CNVs für die HapMap-Proben identifiziert, aber der positive Vorhersagewert sehr niedrig ist. Aufgrund dieser Tatsache haben wir ExomeDepth nicht zum Erfassen von CNVs in den TOF-Patienten verwendet. Im Vergleich zu CoNIFER, haben wir mit unserer Methode mehr CNVs sowohl in den HapMap-Proben als auch in unser TOF-Kohorte gefunden. In der TOF-Kohorte, die acht Fälle umfasst, fanden wir in vier Regionen eine Erhöhung der Kopienzahl in drei Patienten. Alle vier Variationen konnten validiert werden. Darüber hinaus sind darunter drei Gene betroffen, die wichtige Regulatoren in der Herzentwicklung sind (NOTCH1, ISL1) oder sich in einer Region befinden, die bereits mit Herzfehlern assoziiert ist (PRODH).

Der zweite Teil der Dissertation fokussiert sich auf die stabilen Anreicherungsmuster von Histonmodifikationen (H3K4me2 und H3K4me3) in Kombination mit einem gewebespezifischen Transkriptionsfaktor (ChIP-seq von MyoD), der die Muskeldifferenzierung reguliert. Dabei fanden wir spezifische H3K4me2/3 Profile bei muskelrelevanten Genen. Das durchschnittliche H3K4me3-Profil ist im Allgemeinen unmittelbar hinter dem Transkriptionsstart angereichet, während sich H3K4me2 mehr über das ganze Gen verteilt. Darüber hinaus zeigte unsere Studie eine deutlich stärkere Bindung von MyoD an einer besonderen Untergruppe von Genen, die mit einer vorwiegend repressiven Rolle von MyoD einhergeht. Interessanterweise deuten die Ergebnisse daraufhin, dass MyoD während der Muskeldifferenzierung an Patz1 bindet und dieses herunterreguliert, was möglicherweise einen wichtigen Regulationsmechanismus bei der Muskeldifferenzierung aufzeigt.

Schließlich wurde drittens eine Pipeline zur Identifizierung von unterschiedlich verwendeten Exons (ausgeschlossen oder enthaltenen) in RNA-seq Daten entwickelt. Vor mehr als zehn Jahren wurde Dpf3 (auch bekannt als Baf45c) in der AG Sperling als Chomatin-Remodeling-Faktor identifiziert, dessen Expression im rechten Ventrikel von TOF-Patienten signifikant hochreguliert ist. Es wurde gezeigt, dass Dpf3 speziell im Herzen und in den Somiten exprimiert wird und methylierte und acetylierte Lysinreste der Histone 3 und 4 bindet. Darüber hinaus ist bekannt, dass mehrere Proteine, die Chromatin-Histon-Modifikationen binden, mit Spleißfaktoren interagieren. Um die Rolle von Dpf3 beim Spleißen zu analysieren, haben wir die Genexpressionsprofile (mRNA-seq) aus der rechten und linken Herzkammer sowie aus dem Skelettmuskel von Dpf3 Knockout- und Wildtyp-Mäusen miteinander verglichen. Die dabei etablierte Pipeline für die Identifizierung der unterschiedliche Verwendung von Exons basiert grundsätzlich auf der Schätzung des PSI (engl. *percent-spliced-in*). Die Ergebisse zeigten, dass Dpf3 wahrscheinlich keine bedeutende Rolle beim Spleißen spielt; allerdings sind hierzu weitere Untersuchungen erforderlich.

Zusammenfassend habe ich in dieser Dissertation verschiedene computerbasierte Methoden zur Analyse von CNVs in kleinen Patientenkohorten, Mustern von Histonmodifikationen und hinsichtlich der unterschiedlichen Verwendung von Exons entwickelt und angewendet.

Summary

The advent of the high-throughput sequencing (HTS) technology has greatly accelerated research in life sciences. Due to its low cost and high efficiency, it is nowadays commonly used to answer various biological questions. In general, in HTS, the sequence of millions of DNA fragments is determined in parallel and these fragments can in turn be generated using different sequencing methods. With the rapid advancement of HTS technologies, their applications seem almost endless, for example it is now possible to sequence an entire genome in less than one day. Besides whole genome sequencing, HTS has various other applications like targeted resequencing, quantification of gene expression profiles (RNA-seq) and genome-wide identification of protein-DNA interactions such as transcription factor binding sites or chromatin histone marks (ChIP-seq). However, the analysis of the massive datasets generated by HTS is only possible with sophisticated bioinformatics methods. In this thesis, I have presented computational approaches for analyzing data obtained by targeted DNA resequencing, RNA-seq and ChIP-seq, aimed at answering biological questions regarding cardiac disease and skeletal muscle development.

First, a novel copy number variation (CNV) calling method was developed to identify individual disease-relevant CNVs using exome or targeted resequencing data of small sets of samples. Detecting CNVs from targeted resequencing data is difficult due to non-uniform read-depth between captured regions. Moreover, a method was needed to detect personalized CNVs from small cohort of patients without using controls. Thus, we developed such a method and evaluated it using publicly available data of eight HapMap samples, and subsequently applied it to a small number of Tetralogy of Fallot (TOF) patients. In addition to our method, we used the two publicly available tools, namely ExomeDepth and CoNIFER. ExomeDepth identified more CNVs for HapMap samples as compared to CoNIFER and our method; however, the positive predictive value was very low. Therefore, we decided not to use ExomeDepth for detecting CNVs in the TOF patients. Compared to CoNIFER, we identified more CNVs in both the HapMap samples as well as in our TOF cohort. In

the TOF cohort (comprising eight cases), we found four copy number gains in three patients. All four gains could be validated and, in addition, the three genes affected by CNVs were found to be important regulators of heart development (*NOTCH1, ISL1*) or were located in a region already associated with cardiac malformations (*PRODH*).

The second study presented in this thesis was focused on the stable enrichment patterns of histone modifications (H3K4me2 and H3K4me3) in combination with a tissue-specific transcription factor (MyoD) that regulate myogenic differentiation. Here, we found specific H3K4me2/3 profiles on muscle-relevant genes. In general, the average profile of H3K4me3 was enriched directly downstream of transcription start sites, whereas H3K4me2 was located further over the gene body. Furthermore, our study revealed a significant stronger binding of MyoD to this particular subset of genes, with a predominantly repressive role of MyoD. Interestingly, the results suggested that MyoD binds and down-regulates *Patz1* during myogenic differentiation, which might provide an important regulatory mechanism to promote myogenic differentiation.

Finally, a pipeline was developed to identify differential exon usage from RNA-seq data, with the intention of identifying the exons that are excluded or included. Almost a decade ago, the Sperling lab identified Dpf3 (also known as Baf45c) as chromatin remodeling factor, whose expression was significantly up-regulated in the right ventricle of TOF patients. It was shown that *Dpf3* is specifically expressed in heart and somites and binds methylated and acetylated lysine residues of histone 3 and 4. Moreover, it is known that several proteins, which bind chromatin histone modifications, interact with splicing factors. Thus, to dissect the role of Dpf3 in splicing, we compared gene expression profiles (mRNA-seq) generated from the right and left ventricle as well as skeletal muscle of Dpf3 knockout and wild-type mice. Basically, the established pipeline for the identification of the differential exon usage is based on the estimation of percent-spliced-in (PSI, Ψ). The results suggested that Dpf3 might not play a significant role in splicing; however, further investigations are required.

In summary, within this thesis, I have developed and applied different computational methods for analyzing CNVs in small cohorts of patients, patterns of histone modifications and differential exon usage.

Appendix A

Outlier-Based CNV Calling Method

Chr	Start position (hg19)	End position (hg19)	Type of variation	HapMap sample
chr1	155,234,407	155,237,870	gain	NA15510
chr1	155,253,768	155,261,736	gain	NA15510
chr2	240,981,511	240,982,011	gain	NA12878
chr3	19,559,462	19,924,248	gain	NA15510
chr3	20,164,156	20,181,845	gain	NA15510
chr3	20,215,780	20,216,280	gain	NA15510
chr4	68,795,606	68,925,183	gain	NA18517
chr4	68,928,187	68,928,787	gain	NA18517
chr4	68,930,393	68,934,496	gain	NA18517
chr5	69,717,189	69,718,089	gain	NA18517
chr5	69,729,631	69,730,131	gain	NA18517
chr5	70,308,153	70,308,653	gain	NA18517
chr7	99,564,684	99,621,311	gain	NA15510
chr9	108,456,919	108,536,213	gain	NA15510
chr9	117,087,073	117,092,300	gain	NA15510
chr9	40,773,663	40,774,263	gain	NA12878
chr9	41,590,682	41,592,182	gain	NA12878
chr11	4,967,401	4,968,201	gain	NA19240
chr11	5,878,066	5,878,966	loss	NA19240
chr11	6,190,624	6,191,524	loss	NA19129
chr12	133,721,045	133,733,489	gain	NA19240
chr12	133,764,519	133,768,587	gain	NA19240
chr12	133,778,781	133,779,381	gain	NA19240
chr14	106,539,004	106,539,504	gain	NA19240
chr14	106,780,499	106,781,099	gain	NA19240
chr14	21,359,867	21,423,999	loss	NA19240
chr16	21,623,981	21,636,326	gain	NA18517
chr16	21,658,494	21,666,721	gain	NA18517
chr16	21,702,877	21,712,336	gain	NA18517
chr16	21,734,219	21,739,705	gain	NA18517
chr17	39,535,858	39,538,575	gain	NA19240
chr17	44,171,932	44,249,515	gain	NA12878
chr19	43,688,932	43,698,720	gain	NA18517
chr19	9,868,776	9,869,276	loss	NA19129
chr22	20,457,890	20,459,090	gain	NA19129
chr22	21,742,009	21,742,909	gain	NA19129
chr22	21,828,820	21,829,620	gain	NA19129
chr22	21,900,797	21,901,297	gain	NA19129
chr22	22,453,213	22,453,713	loss	NA12878
chr22	23,134,983	23,135,483	loss	NA12878

Table A.1: CNVs found in the five HapMap samples using type10 Dixon's Q test in the outlier-based CNV calling method

Chr	Start position (hg19)	End position (hg19)	Type of variation	HapMap sample
chr1	152,573,211	152,586,435	loss	NA15510
chr1	152,573,211	152,586,435	loss	NA19129
chr1	155,234,407	155,237,870	gain	NA15510
chr1	155,253,768	155,261,736	gain	NA15510
chr2	240,981,511	240,982,311	gain	NA12878
chr3	19,559,462	19,930,107	gain	NA15510
chr3	20,164,156	20,187,926	gain	NA15510
chr3	20,215,780	20,216,280	gain	NA15510
chr4	68,795,606	68,925,183	gain	NA18517
chr4	68,928,187	68,928,787	gain	NA18517
chr4	68,930,393	68,934,496	gain	NA18517
chr4	70,146,232	70,146,832	loss	NA12878
chr4	70,146,232	70,146,932	loss	NA19129
chr4	70,152,473	70,160,559	loss	NA12878
chr4	70,152,473	70,160,559	loss	NA19129
chr5	69,717,189	69,718,089	gain	NA18517
chr5	69,729,631	69,730,131	gain	NA18517
chr5	69,733,151	69,733,651	gain	NA18517
chr5	70,308,153	70,308,753	gain	NA18517
chr7	141,755,347	141,758,103	loss	NA12878
chr7	75,045,612	75,046,112	gain	NA19129
chr7	99,564,684	99,621,311	gain	NA15510
chr9	108,456,919	108,536,213	gain	NA15510
chr9	117,087,073	117,092,300	gain	NA15510
chr9	40,773,663	40,774,263	gain	NA12878
chr9	41,590,682	41,592,182	gain	NA12878
chr11	4,967,401	4,968,301	gain	NA19240
chr11	5,878,066	5,878,966	loss	NA19240
chr11	6,190,624	6,191,524	loss	NA19129
chr11	7,817,616	7,818,416	loss	NA19129
chr11	7,817,616	7,818,416	loss	NA19240
chr12	133,721,045	133,733,489	gain	NA19240
chr12	133,764,519	133,768,587	gain	NA19240
chr12	133,778,781	133,779,381	gain	NA19240
chr14	105,417,358	105,418,158	loss	NA12878
chr14	105,417,358	105,418,158	loss	NA19129
chr14	106,539,004	106,539,504	gain	NA19240
chr14	106,780,499	106,781,099	gain	NA19240
chr14	21,359,867	21,423,999	loss	NA19240
chr15	22,368,674	22,369,374	gain	NA15510
chr15	22,368,674	22,369,374	gain	NA19240
chr15	22,466,012	22,466,512	gain	NA15510
chr15	22,466,012	22,466,512	gain	NA19240
chr15	22,489,704	22,490,204	gain	NA15510
chr16	21,623,981	21,636,326	gain	NA18517
chr16	21,658,494	21,666,721	gain	NA18517
chr16	21,702,877	21,712,336	gain	NA18517
chr16	21,734,219	21,739,705	gain	NA18517
chr16	72,107,785	72,110,923	gain	NA18517
chr16	72,107,785	72,110,923	gain	NA19240
chr17	39,535,858	39,538,575	gain	NA19240
chr17	44,171,932	44,249,515	gain	NA12878
chr19	43,688,932	43,698,720	gain	NA18517
chr19	9,868,176	9,869,276	loss	NA19129
chr22	20,456,590	20,457,090	gain	NA19129
chr22	20,457,690	20,459,090	gain	NA19129
chr22	21,739,909	21,740,409	gain	NA19129
chr22	21,742,009	21,743,009	gain	NA19129
chr22	21,828,820	21,829,620	gain	NA19129
chr22	21,830,142	21,831,242	gain	NA19129
chr22	21,832,798	21,834,188	gain	NA19129
chr22	21,841,563	21,842,863	gain	NA19129
chr22	21,900,797	21,901,397	gain	NA19129
chr22	22,453,213	22,453,713	loss	NA12878
chr22	23,134,983	23,135,483	loss	NA12878

Table A.2: CNVs found in the five HapMap samples using type20 Dixon's Q test in the outlier-based CNV calling method

Listing A.1: R script for our CNV calling method

```
#### Author - Vikas Bansal
#### Email - vikas.bansal@charite.de
#### Created - October 2013
#### R 2.15.1
#### Script S1

library("outliers")
library("HMM")

##-----------------------------
## modified code for Dixon's Q test from "outliers" package, which returns sample names, p-values and outlier type (gain, loss or normal)
##


my.dixon.test <- function (x, type = 0, opposite = FALSE, two.sided = TRUE) {
  DNAME <- deparse(substitute(x))
  x <- sort(x[complete.cases(x)])
  n <- length(x)
  if ((type == 10 || type == 0) & (n < 3 || n > 30))
    stop("Sample size must be in range 3-30 for type10")
  if (type == 20 & (n < 4 || n > 30))
    stop("Sample size must be in range 4-30 for type20")
  if (xor(((x[n] - mean(x)) < (mean(x) - x[1])), opposite)) {
    alt = paste("lowest value", x[1], "is an outlier")
    number="Loss"
    if (type == 10) {
    Q = (x[2] - x[1])/(x[n] - x[1])
    out.patient=names(x[1])
    }
    else {
    Q = (x[3] - x[1])/(x[n] - x[1])
    out.patient=paste(names(x[1]),names(x[2]),sep=";")
    }
  }
  else {
    alt = paste("highest value", x[n], "is an outlier")
    number="Gain"
    if (type == 10) {
    Q = (x[n] - x[n - 1])/(x[n] - x[1])
    out.patient=names(x[n])
```

```
41      }
42      else {
43      Q = (x[n] - x[n - 2])/(x[n] - x[1])
44      out.patient=paste(names(x[n]),names(x[n-1]),sep=";")
45      }
46    }
47    pval <- pdixon(Q, n, type)
48    if (two.sided) {
49      pval <- 2 * pval
50      if (pval > 1)
51      pval <- 2 - pval
52    }
53    RVAL <- list(statistic = c(Q = Q), alternative = alt, p.
          value = pval,
54      method = "Dixon test for outliers", data.name = DNAME, x
          =out.patient, num=number)
55    class(RVAL) <- "htest"
56    return(RVAL)
57  }
58
59  ##---------------
60  ## main function - calling CNVs
61  ## input data frame contains first 4 columns - CHROM, START,
      END, GC% and 5th, 6th, 7th, ... 34th column contains copy
      number value for each sample
62  ## above input data frame can be created from the output of
      mrCaNaVar "out_prefix.copynumber.bed" output file (first
      step of the method)
63  ##
64
65  exomeCNA <- function(df.var, type = 0, w.size = 100, p.cutoff
        = 0.01, two.sided = FALSE, conti.win = 5 ) {
66    col <- ncol(df.var)
67    if (type == 0) {
68      if (col < 12 & col >6) {
69        type <- 10
70      }
71      else if (col < 35 & col >11){
72        type <- 20
73      }
74      else {
75        stop("Sample size must be in range 3-30")
76      }
77    }
78    else if (type != 10 && type != 20) {
```

```
79      stop("Type should be 10 or 20")
80    }
81
82    ## read in the data frame
83    all <- df.var
84    colnames(all)[1:3] <- c("CHROM","START","END")
85    end.all <- df.var[apply(df.var[,5:col],1,function(v)sum(v!=
          0,na.rm=TRUE)>=((col-4)/2)),]
86    colnames(end.all)[1:3] <- c("CHROM","START","END")
87    not.same <- apply(end.all[,5:col],1,function(i) length (
          unique(i)) > 1 )
88    end.all <- end.all[not.same,]
89    one <- col+1
90    two <- col+2
91    three <- col+3
92
93    ## apply type20 Dixon test if type is equal to 20 (second
          step of the method)
94    if (type == 20) {
95      for  (chak in c(10,20)) {
96        ko <- apply(end.all[,5:col],1, function(test){
97          to <- my.dixon.test(test, type=chak ,two.sided= two.
                sided)
98        })
99        end.all[,one] <- sapply(ko,function(la){la$p.value})
100       end.all[,two] <- sapply(ko,function(la){la$x})
101       end.all[,three] <- sapply(ko,function(la){la$num})
102       colnames(end.all)[one:three] <- c(paste("p.value,type",
              chak,sep=""), paste("patients.type",chak,sep=""),
              paste("copynum.type",chak,sep=""))
103       one <- one+3
104       two <- two+3
105       three <- three+3
106     }
107
108     ## return the outlying windows which has p-value less
            than p.cutoff
109     filtered <- (end.all[which(end.all[,col+1] <= p.cutoff |
            end.all[,col+4] <= p.cutoff ),])
110     if(length(filtered)==0 || nrow(filtered) == 0 ){
111       stop("No significant regions found")
112     }
113     else{
114       filtered[which(filtered[,col+1] <= p.cutoff),ncol(
              filtered)+1] <- "type10"
```

```
115         filtered[is.na(filtered)]<- "type20"
116         colnames(filtered)[ncol(filtered)] <- "No. of patients"
117         filtered <- (filtered[which(filtered[,3]-filtered[,2]
              == w.size),])
118         if (length(unique(filtered[,col+3]) ) > 1){
119           filtergain <- (filtered[which(filtered[,col+3]== "
                Gain"),])
120           filterloss <- (filtered[which(filtered[,col+3]!= "
                Gain"),])
121           Patient1 <- vector()
122           Patient2 <- vector()
123           for (chak in 1:nrow(filtergain)){
124             if(filtergain[chak,ncol(filtergain)] == "type10"){
125               Patient1[chak] <- filtergain[chak,col+2]
126               Patient2[chak] <- "NA"
127             }
128             else if (filtergain[chak,ncol(filtergain)] == "
                  type20"){
129               test <- unlist(strsplit(filtergain[chak,col+5],";
                    "))
130               Patient1[chak] <- test[1]
131               Patient2[chak] <- test[2]
132             }
133           }
134           gain <- filtergain[,c(1,2,3)]
135           gain[,4:5] <- c(Patient1,Patient2)
136           colnames(gain)[4:5] <- c("Patient1","Patient2")
137           Patient1 <- vector()
138           Patient2 <- vector()
139           for (chak in 1:nrow(filterloss)){
140             if(filterloss[chak,ncol(filterloss)] == "type10"){
141               Patient1[chak] <- filterloss[chak,col+2]
142               Patient2[chak] <- "NA"
143             }
144             else if (filterloss[chak,ncol(filterloss)] == "
                  type20"){
145               test <- unlist(strsplit(filterloss[chak,col+5],";
                    "))
146               Patient1[chak] <- test[1]
147               Patient2[chak] <- test[2]
148             }
149           }
150           loss <- filterloss[,c(1,2,3)]
151           loss[,4:5] <- c(Patient1,Patient2)
152           colnames(loss)[4:5] <- c("Patient1","Patient2")
```

```
153         }
154         else if(unique(filtered[,col+3])[1] == "Gain") {
155           filtergain <- (filtered[which(filtered[,col+3]== "
                Gain"),])
156           Patient1 <- vector()
157           Patient2 <- vector()
158           for (chak in 1:nrow(filtergain)){
159             if(filtergain[chak,ncol(filtergain)] == "type10"){
160               Patient1[chak] <- filtergain[chak,col+2]
161               Patient2[chak] <- "NA"
162             }
163             else if (filtergain[chak,ncol(filtergain)] == "
                  type20"){
164               test <- unlist(strsplit(filtergain[chak,col+5],";
                    "))
165               Patient1[chak] <- test[1]
166               Patient2[chak] <- test[2]
167             }
168           }
169           gain <- filtergain[,c(1,2,3)]
170           gain[,4:5] <- c(Patient1,Patient2)
171           colnames(gain)[4:5] <- c("Patient1","Patient2")
172         }
173         else {
174           filterloss <- (filtered[which(filtered[,col+3]!= "
                Gain"),])
175           Patient1 <- vector()
176           Patient2 <- vector()
177           for (chak in 1:nrow(filterloss)){
178             if(filterloss[chak,ncol(filterloss)] == "type10"){
179               Patient1[chak] <- filterloss[chak,col+2]
180               Patient2[chak] <- "NA"
181             }
182             else if (filterloss[chak,ncol(filterloss)] == "
                  type20"){
183               test <- unlist(strsplit(filterloss[chak,col+5],";
                    "))
184               Patient1[chak] <- test[1]
185               Patient2[chak] <- test[2]
186             }
187           }
188           loss <- filterloss[,c(1,2,3)]
189           loss[,4:5] <- c(Patient1,Patient2)
190           colnames(loss)[4:5] <- c("Patient1","Patient2")
191         }
```

```
192     }
193   }
194
195   ## apply type10 Dixon test if type is equal to 10 (second
          step of the method)
196   else{
197     chak=10
198     ko <- apply(end.all[,5:col],1, function(test){
199       to <- my.dixon.test(test, type=chak ,two.sided= two.
              sided)
200     })
201     end.all[,one] <- sapply(ko,function(la){la$p.value})
202     end.all[,two] <- sapply(ko,function(la){la$x})
203     end.all[,three] <- sapply(ko,function(la){la$num})
204     colnames(end.all)[one:three] <- c(paste("p.value,type",
          chak,sep=""), paste("patients.type",chak,sep=""),
          paste("copynum.type",chak,sep=""))
205     filtered <- (end.all[which(end.all[,col+1] <= p.cutoff )
          ,])
206     if(length(filtered)==0 || nrow(filtered) == 0 ){
207       stop("No significant regions found")
208     }
209     else{
210       filtered[,ncol(filtered)+1] <- "type10"
211       colnames(filtered)[ncol(filtered)] <- "No. of patients"
212       filtered <- (filtered[which(filtered[,3]-filtered[,2]
            == w.size),])
213       if (length(unique(filtered[,col+3]) ) > 1){
214         filtergain <- (filtered[which(filtered[,col+3]== "
              Gain"),])
215         filterloss <- (filtered[which(filtered[,col+3]!= "
              Gain"),])
216         gain <- filtergain[,c(1,2,3, col+2, 4)]
217         loss <- filterloss[,c(1,2,3, col+2, 4)]
218         colnames(gain)[4:5] <- c("Patient1","Patient2")
219         colnames(loss)[4:5] <- c("Patient1","Patient2")
220       }
221       else if(unique(filtered[,col+3])[1] == "Gain") {
222         filtergain <- (filtered[which(filtered[,col+3]== "
              Gain"),])
223         gain <- filtergain[,c(1,2,3, col+2, 4)]
224         colnames(gain)[4:5] <- c("Patient1","Patient2")
225       }
226       else {
227         filterloss <- (filtered[which(filtered[,col+3]!= "
```

```
            Gain") ,])
228         loss <- filterloss[,c(1,2,3, col+2, 4)]
229         colnames(loss)[4:5] <- c("Patient1","Patient2")
230       }
231     }
232   }
233
234   ## apply HMM for each sample separately (third step of the
        method)
235   pat.id <- colnames(end.all)[5:col]
236   for(file in pat.id){
237     if(exists("gain")){
238       gain.sff <- gain[which(gain[,4] == file | gain[,5] ==
            file ),1:5]
239     }
240     else {
241       gain.sff <- data.frame(a=character(0))
242     }
243     if(exists("loss")){
244       loss.sff <- loss[which(loss[,4] == file | loss[,5] ==
            file  ),1:5]
245     }
246     else {
247       loss.sff <- data.frame(a=character(0))
248     }
249     all.win <- all[,1:4]
250     if(length(gain.sff)==0 || nrow(gain.sff) == 0 ){
251        if(length(loss.sff)==0 || nrow(loss.sff) == 0 ){
252         next
253        }
254        else{
255         loss.sff[,6] <- "loss"
256         lossgain78 <- loss.sff
257        }
258     }
259     else if (length(loss.sff)==0 || nrow(loss.sff) == 0 ) {
260       gain.sff[,6] <- "gain"
261       lossgain78 <- gain.sff
262     }
263     else {
264       loss.sff[,6] <- "loss"
265       gain.sff[,6] <- "gain"
266       lossgain78 <- (rbind(gain.sff,loss.sff))
267     }
268     merge78 <- (merge(all.win,lossgain78,by = c("CHROM","
```

```
        START","END"),all.x=TRUE))
269     merge78[is.na(merge78)] <- "normal"
270     forhmm78 <- merge78[,c(1:3,7)]
271     forhmm78[,5] <- "wait"
272     colnames(forhmm78)[5] <- "After.HMM"
273
274     ## initial transition and emission probabilities
275     hmm <- initHMM(c("gain","loss","normal"), c("gain","loss
            ","normal"), transProbs=matrix(c
            (.6,.2,.2,.2,.6,.2,.2,.2,.6),3),emissionProbs=matrix(c
            (.6,.2,.2,.2,.6,.2,.2,.2,.6),3))
276
277     ## recomputing transition and emission probabilities
            using the Baum-Welch algorithm
278     ## finding most likely sequence of the hidden states by
            the Viterbi algorithm
279     for(jo in unique(forhmm78[,1])){
280       cat("\r", paste(jo,"-",file) , "\n")
281       observations <- forhmm78[forhmm78[,1] == jo ,4]
282       bw <- baumWelch(hmm, observations,10)
283       viterbi <- viterbi(bw$hmm, observations)
284       forhmm78[forhmm78[,1]==jo,5] <- viterbi
285       colnames(forhmm78)[4] <- "Before.HMM"
286     }
287
288     ## calling CNV if 5 continuous windows (default conti.win
            = 5) are present with same copy number type
289     forhmm78=forhmm78[,-4]
290     forhmm78$conseq <-cumsum(c(1, forhmm78$After.HMM[-1] !=
            forhmm78$After.HMM[-length(forhmm78$After.HMM)] ) )
291     final <- do.call( rbind,
292     by(forhmm78, list(forhmm78$CHROM, forhmm78$conseq),
293       function(df)
294       if( NROW(df) >= conti.win & df$After.HMM[1] %in% c("
              gain", "loss") ) {
295         cbind(df[1, c("CHROM", "START")] , df[NROW(df), c("
              END", "After.HMM")] )
296         } else{NULL} ) )
297
298     ## output CNVs for each sample if present
299     if (length(final)==0 || nrow(final) == 0) {
300       next
301     }
302     else {
303       colnames(final)[4] <- "TYPE"
```

```
304         write.table(final, file=paste("hmm.",file,sep=""), sep=
                "\t", quote=FALSE, row.names=FALSE)
305       }
306     }
307 }
```

Appendix B

Analysis of Epigenetic Changes during Myogenic Differentiation

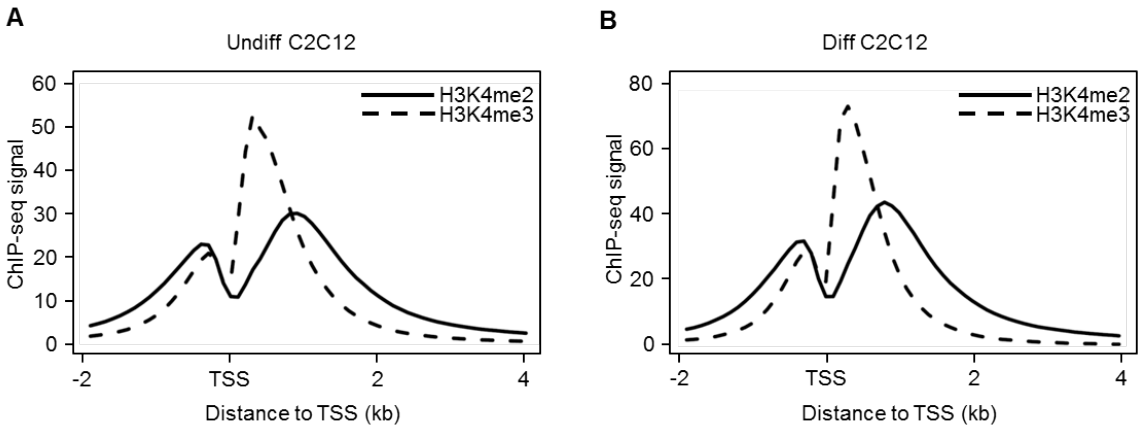


Figure B.1: Average profile of H3K4me2 and H3K4me3 in Undiff and Diff C2C12 cells. (A) Average profile of H3K4me2 and H3K4me3 in Undiff C2C12 and (B) Diff C2C12 cells around the transcription start site (TSS).

	Undiff C2C12		Diff C2C12	
Clusters	H3K4me2/3	H3K4me2/3 + MyoD	H3K4me2/3	H3K4me2/3 + MyoD
1	347	104	362	241
2	895	201	1,007	555
3	1,286	159	1,405	577
4	873	127	943	393
5	1,672	300	1,660	746
6	3,247	547	3,364	1,316

Table B.1: Number of genes in H3K4me2/3 clusters bound by MyoD.

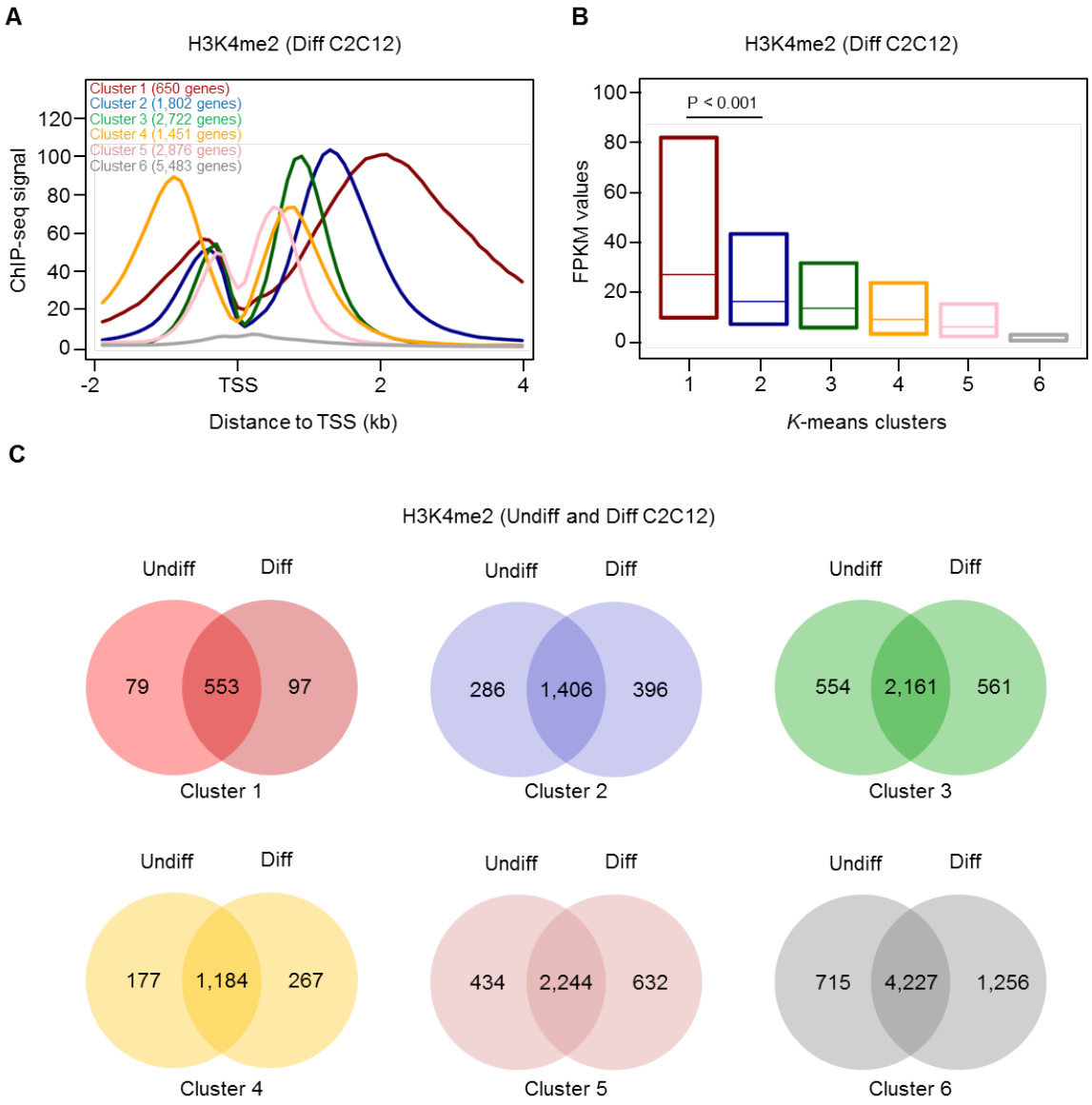


Figure B.2: Clustering analysis of H3K4me2 profiles in differentiated C2C12 cells. (A) H3K4me2 profiles identified by k-means clustering. The clustering is based on the transcription start site (TSS) and the corresponding number of genes is given for each cluster. Genes with multiple TSS can be present in more than one cluster. (B) The box plot (25% to 75% quartile) shows the levels of gene expression (FPKM values) of the different H3K4me2 clusters in Diff C2C12 cells. The expression of cluster 1 and cluster 2 genes was compared using the Mann-Whitney U test. (C) Overlap of genes between the clusters of H3K4me2 in Undiff and Diff C2C12 cells.

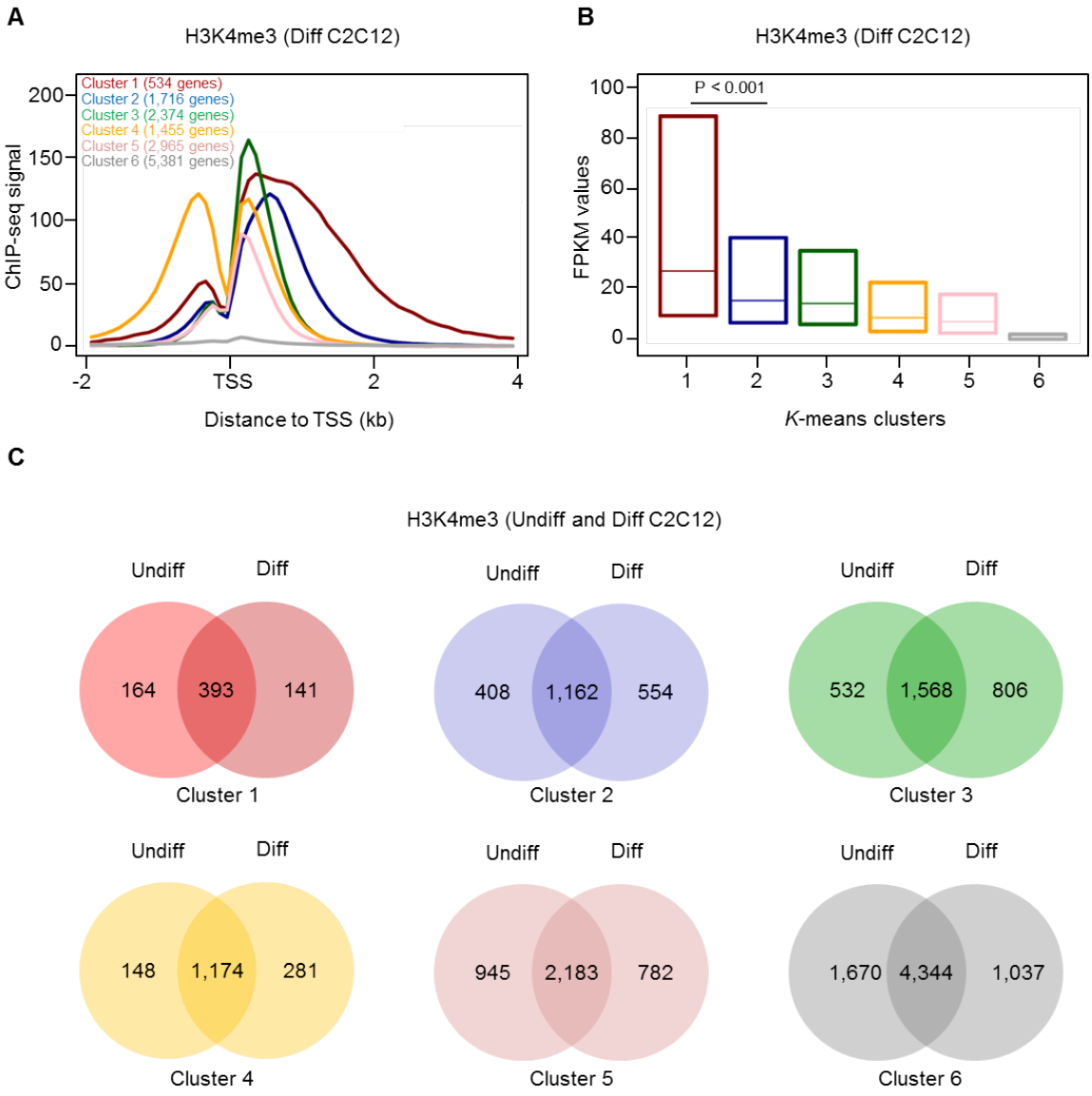


Figure B.3: Clustering analysis of H3K4me3 profiles in differentiated C2C12 cells. (A) H3K4me3 profiles identified by k-means clustering. The clustering is based on the transcription start site (TSS) and the corresponding number of genes is given for each cluster. Genes with multiple TSS can be present in more than one cluster. (B) The box plot (25% to 75% quartile) shows the levels of gene expression (FPKM values) of the different H3K4me3 clusters in Diff C2C12 cells. The expression of cluster 1 and cluster 2 genes was compared using the Mann-Whitney U test. (C) Overlap of genes between the clusters of H3K4me3 in Undiff and Diff C2C12 cells.

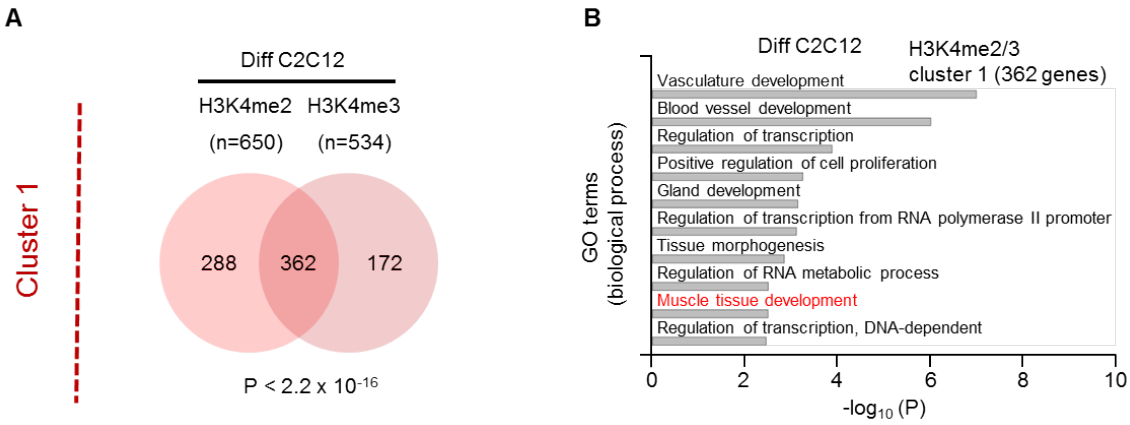


Figure B.4: Comparison of H3K4me2 and H3K4me3 cluster 1 in differentiated C2C12 cells. (A) Overlap of H3K4me2 and H3K4me3 cluster 1 genes in Diff C2C12 cells. The P-value is based on a hypergeometric test. (B) GO enrichment analysis of common cluster 1 genes using the DAVID database. Top ten biological process terms with an adjusted (Benjamini-Hochberg) P-value ≤ 0.01 are indicated. GO terms related to muscle development are highlighted in red.

Listing B.1: Perl script for identifying overrepresented motifs

```
#!/usr/bin/perl

##################

#Usage: perl Over_represented_motifs.pl input.fasta

#COMMENTS: This script finds the overrpresented motifs in
    input sequences as compared to background control.

#Using Binomial test to calculate Pvalues.


##################


use warnings;
use strict;

print "
######################

Over_represented_motifs.pl <- By Vikas Bansal

Usage: perl Over_represented_motifs.pl input.fasta

May 2015

This script finds the overrpresented motifs in input
    sequences.

User can provide the control file or it can be generated
    randomly.

Make sure R is installed.

This script uses MATCH tool provided by Transfac database
    which is installed in /project/archive/Biobase/
    TRANSFAC_matrices/match

Please don't delete that directory :) !!!!!!!!!!
#################################\n\n";


```

```
39 print "
40 ########################
41
42 Did you read the above comments and agree not to delete the
     directory? Y/N
43
44 ############################\n\n";
45
46 my $agree = <STDIN>;
47 chomp $agree;
48 if($agree eq "Y" || $agree eq "y") {
49
50   }else{
51
52   print "\n\nBYE BYE!!!!!!!!!!!!!!!!!!!!\n\n";
53     exit;
54   }
55
56
57
58 my $Filename;
59
60 # Check if input file exists or not
61 if ($ARGV[0]) {
62   $Filename = $ARGV[0];
63 }
64 else {
65   print "Not enough arguments\n\n";
66   die "!\n\n";
67 }
68
69 open (FILE, "<$Filename") or die "\n\nCannot open file
     $Filename!!!!!!!! BYE BYE!!!!!!!!!!!!!!!!!!!!\n\n";
70 close FILE;
71
72 my $control_file_name;
73
74
75
76
77 ## Which profile to use for overrepresented motifs
78 print "
79 ############################
80
81
```

```
82 Select the matrix profile you would like to use. For example,
       if you are working with vertebrates then select
   vertebrate specific.
83
84 1 adipocyte_specific.prf                    2
   nerve_system_specific.prf
85 3 bacteria.prf                             4
   pancreatic_beta_cell_specific.prf
86 5 cell_cycle_specific.prf                  6 pituitary_specific.
   prf
87 7 fungi.prf                                8 plants.prf
88 9 immune_cell_specific.prf                 10 redox_specific.prf
89 11 insects.prf                             12
   vertebrate_non_redundant_minFN.prf
90 13 invertebrates.prf                       14
   vertebrate_non_redundant_minFP.prf
91 15 liver_specific.prf                      16
   vertebrate_non_redundant_minSUM.prf
92 17 lung_specific.prf                       18
   vertebrate_non_redundant.prf
93 19 muscle_specific.prf                     20 vertebrates.prf
94 21 nematodes.prf
95
96
97 ##Comments:
98
99 The search algorithm uses two score values: the matrix
   similarity score (MSS) and the core similarity score (CSS)
100
101 These two scores measure the quality of a match between the
   sequence and the matrix, which ranges from 0.0 to 1.0,
   where 1.0 denotes an exact match
102
103 The core of each matrix is defined as the first five most
   conserved consecutive positions of a matrix. The core
   similarity score is calculated for all pentanucleotides
   and prolonged at both ends, so that it fits the matrix
   length.
104
105 MSS and CSS are precalculated by TRANSFAC database to
   minimize false negatives. In addition, for vertebrates
   they provide three different cut offs for non redundant i.
   e.
106
107   - Cut-off to Minimize False Negative Matches (minFN)
```

```
108   - Cut-off to Minimize False Positive Matches (minFP)
109   - Cut-off to Minimize the Sum of Both Error Rates (minSUM)
110
111 ##############################\n\n";
112
113   my $profile = <STDIN>;
114   chomp $profile;
115
116   if($profile <1 || $profile >21 ) {
117   print "\n\nNumber should be between 1 to 21 \n\nBYE BYE
          !!!!!!!!!!!!!!!!!!!!\n\n";
118     exit;
119
120   }
121
122 my @profile_names=("empty", "adipocyte_specific.prf", "
      nerve_system_specific.prf", "bacteria.prf", "
      pancreatic_beta_cell_specific.prf", "cell_cycle_specific.
      prf", "pituitary_specific.prf", "fungi.prf", "plants.prf",
       "immune_cell_specific.prf" , "redox_specific.prf", "
      insects.prf" , "vertebrate_non_redundant_minFN.prf", "
      invertebrates.prf", "vertebrate_non_redundant_minFP.prf",
      "liver_specific.prf", "vertebrate_non_redundant_minSUM.prf
      ", "lung_specific.prf", "vertebrate_non_redundant.prf", "
      muscle_specific.prf", "vertebrates.prf", "nematodes.prf");
123
124
125
126
127 ##Ask user if comntrol file exists or would like to create
128 print "
129 ############################
130
131 Do you have your own control sequences? Y/N
132
133 ########################\n\n";
134
135   my $control = <STDIN>;
136   chomp $control;
137
138   if($control eq "N" || $control eq "n") {
139     print "\n\n\n Generating random control sequences\n\n\n";
140
141     $control_file_name =random_dna_strings();
142
```

```
143     chomp $control_file_name;
144   }elsif($control eq "Y" || $control eq "y") {
145
146     print "\nPlease, type the name of the control file\n\n";
147                 $control_file_name =<STDIN>;
148     chomp $control_file_name;
149     open (FILE, "<$control_file_name") or die "\n\nCannot
          open file $control_file_name!!!! BYE BYE
          !!!!!!!!!!!!!!!!!!!!\n\n";
150     close FILE;
151   }else{
152
153   print "\n\nPlease write Y or N \n\nBYE BYE
        !!!!!!!!!!!!!!!!!!!!\n\n";
154     exit;
155   }
156
157
158
159
160
161 print "\n\nRunning MATCH and using profile $profile_names[
      $profile]\n\n";
162
163
164 ##Running MATCH from TRANSFAC
165 my $match_cmd= join('','/project/archive/Biobase/
      TRANSFAC_matrices/match/bin/match ', '/project/archive/
      Biobase/TRANSFAC_matrices/match/data/matrix.dat ',
      $control_file_name, ' random_match_out.txt', ' /project/
      archive/Biobase/TRANSFAC_matrices/match/data/prfs/',
      $profile_names[$profile]);
166
167 system ($match_cmd);
168
169 print "\n\nMATCH completed for $control_file_name\n\n";
170
171 my $input_match_cmd= join('','/project/archive/Biobase/
      TRANSFAC_matrices/match/bin/match ', '/project/archive/
      Biobase/TRANSFAC_matrices/match/data/matrix.dat ',
      $Filename, ' input_match_out.txt', ' /project/archive/
      Biobase/TRANSFAC_matrices/match/data/prfs/',$profile_names
      [$profile]);
172
173 system ($input_match_cmd);
```

```
174
175 print "\n\nMATCH completed for $Filename\n\n";
176
177
178
179
180 ##COunt number of sequences has particular motif
181 my $motif_input_temp=`awk 'NR>5{print \$1,\$4}'
    input_match_out.txt    | awk 'NF>0' > input_unique_temp.txt
    `;
182
183
184 my $motif_random_temp=`awk 'NR>5{print \$1,\$4}'
    random_match_out.txt    | awk 'NF>0' > random_unique_temp.
    txt`;
185
186 system ($motif_input_temp);
187 system ($motif_random_temp);
188
189
190
191
192 my $input_file_match = "input_unique_temp.txt";
193 my $input_file_match_out = "input_file_uniq";
194
195 open (FILE, "<$input_file_match") or die "Cannot open file
    $input_file_match !!!!: $!";
196 open (OUT1, ">$input_file_match_out") or die "Cannot open
    file $input_file_match_out !!!!: $!";
197 my ($ps, $pe) = split(/\s/,<FILE>);
198
199 while (<FILE>) {
200 chomp;
201     my ($cs, $ce) = split;
202 if($cs ne "Inspecting"){
203       print OUT1 $cs, "\t",$pe,"\n";
204 }       else{($ps,$pe)=($cs,$ce);}
205
206
207 }
208 close FILE;
209
210
211
212
```

```
213 my $random_file_match = "random_unique_temp.txt";
214 my $random_file_match_out = "random_file_uniq";
215
216 open (FILE, "<$random_file_match") or die "Cannot open file
    $random_file_match !!!!: $!";
217 open (OUT2, ">$random_file_match_out") or die "Cannot open
    file $random_file_match_out !!!!: $!";
218 my ($psr, $per) = split(/\s/,<FILE>);
219
220 while (<FILE>) {
221 chomp;
222     my ($cs, $ce) = split;
223 if($cs ne "Inspecting"){
224       print OUT2 $cs, "\t",$per,"\n";
225 }       else{($psr,$per)=($cs,$ce);}
226
227
228 }
229
230
231 close FILE;
232 close OUT1;
233 close OUT2;
234
235
236
237
238
239 ##Use R for binomial test
240 my $motif_input=`sort -u input_file_uniq | cut -f1 | grep '
    \\\$' | sort | uniq -c > input_motifs.txt`;
241
242
243 my $motif_random=`sort -u random_file_uniq | cut -f1 | grep '
    \\\$' | sort | uniq -c > random_motifs.txt`;
244
245 system ($motif_input);
246 system ($motif_random);
247
248
249 system(`grep "Inspect" input_match_out.txt | wc -l >
    number_of_input_seq.txt`);
250 system(`grep "Inspect" random_match_out.txt | wc -l >
    number_of_random_seq.txt`);
251
```

```
252 print "\n\nCalculating P-values for overrepresented motifs\n\
    n";
253
254 system('Rscript Binomial.R');
255
256 print "\n\nPutting results in overrepresented_motifs_sorted.
    txt\n\n";
257
258 ##Remove temporary files
259 system('rm input_motifs.txt');
260
261 system('rm random_motifs.txt');
262 system('rm number_of_input_seq.txt');
263
264 system('rm number_of_random_seq.txt');
265 system('rm input_file_uniq');
266
267 system('rm input_unique_temp.txt');
268 system('rm random_file_uniq');
269
270 system('rm random_unique_temp.txt');
271
272
273 ###################################
274
275 #random_dna_strings.pl <- PERL SCRIPT WRITTEN BY BENJAMIN
    TOVAR
276
277 #COMMENTS: This script takes arguments given by the user such
     the nucleotide frequencies of each
278
279 #nucleotide (in a scale from 0.0 to 1.0), generates a \"n\"
    number of sequences of \"n\" length
280
281 #with a FASTA header also given by the user and finally
    prints an output file in FASTA format
282
283 ###################################\n\n";
284
285 ##### seed the random stuff #####
286 sub random_dna_strings{
287 srand(time|$$);
288
289 ###### Set the number of iterations (number of random
    sequences to generate) ######
```

```
290
291     print "1) Please type the number of iterations (How many
            random sequences do you want):
292
293     EXAMPLE: \"10\" \n\n";
294
295     my $iterations = <STDIN>;
296     chomp $iterations;
297
298 ###### Set the length of the random DNA strings (how many
        nucleotides length) ######
299
300     print "\n2) Please type the length of the random DNA
            strings (how many nucleotides length):
301
302     EXAMPLE: \"50\" \n\n";
303
304     my $length = <STDIN>;
305     chomp $length;
306
307 ############################ SET THE VALUE OF THE USER'S
        ARGUMENTS #################
308
309 # How much A% content per string:
310
311     print "\n3) Please type the probability distribution of A
            content:
312
313     REMEMBER THAT THE SUM OF THE FOUR PROBABILITIES MUST BE
            EQUAL TO \"1.00\"
314
315     EXAMPLE: \"0.25\" \n\n";
316
317     my $A_content = <STDIN>;
318
319     print "
320     ################################
321     # From a value of \"1.00\" as total probability, there
            are: ", (1-($A_content)),"  available
322     ################################\n\n";
323
324 # How much T% content per string:
325
326     print "\n4) Please type the probability distribution of T
            content:
```

```
327
328     REMEMBER THAT THE SUM OF THE FOUR PROBABILITIES MUST BE
            EQUAL TO \"1.00\"
329
330     EXAMPLE: \"0.25\" \n\n";
331
332     my $T_content = <STDIN>;
333
334     print "
335     ####################################
336     # From a value of \"1.00\" as total probability, there
            are: ", (1-($A_content+$T_content)),"  available
337     ####################################\n\n";
338
339 # How much G% content per string:
340
341     print "\n5) Please type the probability distribution of G
            content:
342
343     REMEMBER THAT THE SUM OF THE FOUR PROBABILITIES MUST BE
            EQUAL TO \"1.00\"
344
345     EXAMPLE: \"0.25\" \n\n";
346
347     my $G_content = <STDIN>;
348
349     print "
350     ####################################
351     # From a value of \"1.00\" as total probability, there
            are: ", my $C_content = (1-($A_content+$T_content+
            $G_content)),"  available
352     ####################################\n\n";
353
354 # How much C% content per string:
355
356     print "\n6) Setting the probability distribution of C
            content\n\n";
357
358     print $C_content,"\n";
359
360 #### Ask the user for the name of the fasta header
361
362     print "\n7) Please, type the name of the fasta header for
            each sequence (is not necessary to put the >):
363
```

```
364     EXAMPLE: \"random_seq\"\n\n";
365
366     my $fasta_header_name =<STDIN>;
367
368 #### Ask the user for the name of output file
369
370     print "\n8) Please, type the name of the output file:
371
372     EXAMPLE: \"random_sequences_set.fa\"\n\n";
373
374     my $output_file_name =<STDIN>;
375
376 ####### ERASE WHITE SPACES OF THE <STDIN> INPUTS #######
377
378     chomp ($A_content,$T_content,$G_content,$C_content,
            $fasta_header_name,$output_file_name);
379
380 ##### Pass the values of the scalar variables to an array
        variable #####
381
382     my @distribution = ($A_content,$T_content,$G_content,
            $C_content);
383
384 ################################## RESULTS SUMMARY
        ##################################
385
386   print "
387 ---------------------------------- RESULTS SUMMARY
        ----------------------------------
388
389 SUCCESS: Here is the $iterations iterations of $length
        nucleotides length of
390 DNA strings in FASTA format with probabilities of:
391
392 A = $A_content
393 T = $T_content
394 C = $G_content
395 G = $C_content
396
397 EXPORTED TO FHE FILE: \"$output_file_name\"
398
399 ------------------------------------------------\n\n";
400
401 ############ OUTPUT FILE SETTINGS #######
402
```

```
403 # Name of the output file
404
405     my $output_file = "$output_file_name";
406     #return("$output_file_name");
407 # Set the file handle "OUTPUT".
408
409     open (OUTPUT_SEQ, ">$output_file");
410
411 ################################## PROGRAM'S MAIN ENGINE
    ##############################
412
413
414 for(my $k=0;$k<$iterations;$k++){
415   print OUTPUT_SEQ ">",$fasta_header_name,"_",($k+1),"\n" if(
        $k==0);
416   print OUTPUT_SEQ "\n>",$fasta_header_name,"_",($k+1),"\n"
        if($k>0);
417
418     for(my $i=0;$i<$length;$i++){
419
420       print OUTPUT_SEQ distribution(@distribution);
421     }
422 }
423 return("$output_file_name");
424 }
425
426
427
428 ############################
429 # distribution
430 # A subroutine to generate random strings depending on the
    probability distribution
431 # of each nucleotide taken from James Tisdall's Beginning
    Perl for Bioinformatics
432 ############################
433
434 sub distribution{
435
436   my @probability = @_;
437
438   unless ($probability[0] + $probability[1] + $probability[2]
        + $probability[3] == 1){
439
440     print "Sum of probabilites must be equal to \"1.0\"!\n";
441     exit;
```

```
442         }
443   my $randnum = rand(1);
444
445   if($randnum < $probability[0]) {
446     return 'A';
447   }elsif($randnum < $probability[0] + $probability[1]) {
448     return 'T';
449   }elsif($randnum < $probability[0] + $probability[1] +
        $probability[2]) {
450     return 'C';
451   }else{
452     return 'G';
453   }
454
455 }
```

Listing B.2: R script for Binomial test for overrepresented motifs

```
1  #R
2  motif_our=(read.table("input_motifs.txt", stringsAsFactors=F,
       header=F))
3  motif_random=(read.table("random_motifs.txt",
       stringsAsFactors=F, header=F))
4  number_of_seq=read.delim("number_of_input_seq.txt",
       stringsAsFactors=F,header=F)
5  number_of_random_seq=read.delim("number_of_random_seq.txt",
       stringsAsFactors=F,header=F)
6  merge_motif=(merge(motif_our, motif_random, by="V2",all=F))
7
8  ##probability of success from random sequences
9  merge_motif[,4]=(merge_motif[,3]/number_of_random_seq[1,1])
10
11 merge_motif$V4[(merge_motif[,4]>1)]=1
12
13
14
15 merge_motif[,5]=(apply(merge_motif[,c(2,4)],1,function(x)
       binom.test(x[1],number_of_seq[1,1]
16 ,p=x[2],alternative="greater")$p.value))
17
18
19
20
21 merge_motif[,6]=(p.adjust(merge_motif[,5],method="bonferroni"
       ))
```

```
22
23 final_motifs_sorted=(merge_motif[(order(merge_motif[,6])),])
24
25
26 colnames(final_motifs_sorted)[1:6]=c("Matrix_name","Num_Input
     _sequences","Num_Control_sequences","Binomial_success","P_
     value","P_adjust_bonferroni")
27
28
29 agg_tf_name=read.delim("/project/archive/Biobase/TRANSFAC_
     matrices/match/data/matrix_tf_name_vikas_final.txt",
     stringsAsFactors=F, header=F)
30
31 final_motifs_sorted[,7]=agg_tf_name[(match(final_motifs_
     sorted[,1],agg_tf_name[,1])),2]
32 colnames(final_motifs_sorted)[7]=("TF_names")
33
34 write.table(final_motifs_sorted,file="overrepresented_motifs_
     sorted.txt",sep="\t",quote=F,row.names=F)
```

Curriculum Vitae

"For reasons of data protection, the curriculum vitae is not published in the electronic version"

"For reasons of data protection, the curriculum vitae is not published in the electronic version"

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst 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, February 2016