Bioinformatics Studies on the Mechanisms of Gene Regulation in Vertebrates

Awaisa Ghazal
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics
Uppsala

Doctoral Thesis
Swedish University of Agricultural Sciences
Uppsala 2014

Acta Universitatis agriculturae Sueciae
2014:82

Cover: Bioinformatics studies on the mechanisms of gene regulation in human, mouse and chicken.
(photo: Awaisa Ghazal)

ISSN 1652-6880
ISBN (print version) 978-91-576-8112-6
ISBN (electronic version) 978-91-576-8113-3

Print: SLU Service/Repro, Uppsala 2014

Bioinformatics Studies on the Mechanisms of Gene Regulation in Vertebrates

Abstract

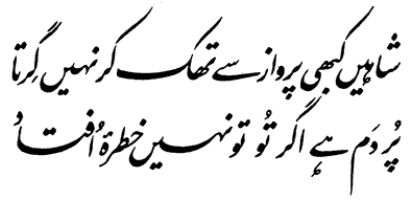
Gene regulation implicates an important role in organismal complexity through contributing to phenotypic variation. In my thesis, I studied three diverse components of gene regulatory networks including transcription factors, DNA methylation and microRNAs to elucidate their role in vertebrates. The first three studies concern a recently identified transcription factor called ZBED6 and the fourth study is based on analysis of chicken miRNAs. First study focused on phylogenetic reconstruction of the *ZBED* gene family, which is comprised of *ZBED1*, *ZBED2*, *ZBED3*, *ZBED4*, *ZBED5* and *ZBED6*. This study demonstrated that ZBEDs arose from at least two independent transposon domestication events in jawed vertebrate ancestors and also identified a new member of this gene family named *C7ORF29*. Second study focused on the role of ZBED6 in colon adenocarcinoma. We scanned DNA methylation and RNA sequencing data generated from colon adenocarcinoma tumor and normal tissues. This analysis unraveled a list of potential ZBED6 binding sites that were hypermethylated and showed a negative correlation with gene expression data. This candidate list contained several genes with a known role in cancer development and showed that in addition to the ZBED6 binding sites another transcription factor PAX5 binding sites were located in its close vicinity. In the third study, functional analyses of ZBED6 were performed. This study revealed that ZBED6 regulates muscle proteins by directly regulating *IGF2* and *Twist2* transcription. Analysis of histone modifications showed that ZBED6 preferentially binds with active promoters to modulate transcriptional activity rather than recruiting repressive histone marks. In the fourth study, potential role of miRNAs in differential growth of two extremely divergent chicken lines was investigated. Our analysis revealed the expression of both known and novel miRNAs in three tissues hypothalamus, liver and pectoralis major muscle of chickens. This study provides a detailed insight into the miRNA expression pattern and suggests an association of miRNAs with growth in chickens. These studies on diverse components will add valuable knowledge to existing information in gene regulatory networks and facilitate in future investigations to reveal regulatory mechanisms underlying diverse biological processes.

Keywords: Bioinformatics, ZBED6, DNA methylation, Colon adenocarcinoma, miRNA, PAX5, Gene regulation

Author's address: Awaisa Ghazal, SLU, Department of Animal Breeding and Genetics, P.O. Box 7023, 750 07 Uppsala, Sweden
E-mail: Awaisa.Ghazal@ slu.se

Dedication

To my family, teachers and friends.



The hawk is never tired of flight, does not drop gasping on the ground:
If unwearied it remains on wings, from hunters′ dread is safe and sound.

Allama Iqbal

Contents

List of Publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I Hayward, A*., **Ghazal, A***., Andersson, G., Andersson, L., and Jern, P. (2013). ZBED evolution: repeated utilization as regulators of diverse host functions. PLoS ONE 8(3): e59940. (*AH and AG contributed equally)

II **Ghazal, A**., Wallerman, O., Johansson, A., Andersson, L., Andersson, G., and Molin, A-M. Genome-wide analysis of the DNA methylation landscape in colon adenocarcinoma reveals hyper-methylation at ZBED6-target regions. Manuscript.

III Jiang, L., Wallerman, O., Younis, S., Rubin, C.J., Gilbert, E.R., Sundstrom, E., **Ghazal, A**., Zhang, X., Wang, L., Mikkelsen, T.S., Andersson, G. & Andersson, L. (2014). ZBED6 modulates the transcription of myogenic genes in mouse myoblast cells. *PLoS One,* 9(4), p. e94187.

IV **Ghazal, A**., Molin, A-M., Lopes Pinto, FA., De Koning, D-J., Andersson, L., and Andersson, G. Differentially expressed microRNAs in relation to growth of two extremely divergent chicken lines: LWS and HWS. Manuscript.

The contribution of Awasia Ghazal to the papers included in this thesis was as follows:

I Actively involved in planning and phylogenetic analysis and wrote first draft of manuscript

II Participated in planning, analysis of DNA methylation and RNA sequencing data and writing of manuscript

III Contributed in transcription factor binding site analysis and limited participation in writing

IV Involved in planning and performing of miRNA transcriptome analysis and contributed in writing of manuscript.

Abbreviations

ChIP	Chromatin immunoprecipitation
CIMP	CpG island methylator phenotype
DE	Differentially expressed
DHss	DNase I hypersensitive site
DMRs	Differentially methylated regions
DNMT	DNA methyltransferase
EMSA	Electrophoretic mobility shift assay
ENCODE	The encyclopaedia of DNA elements
HATs	Histone acetylase
HDACs	Histone deacetylase
HWS	High weight selection lines
IGF2	Insulin-like growth factor II
Inr	Initiator
LWS	Low weight selection lines
miRNA	MicroRNA
myr	Million years
NGS	Next generation sequencing
piRNAs	Piwi-interacting RNAs
POL	Polymerase
QTL	Quantitative trait locus
QTN	Quantitative trait nucleotide
RISC	RNA induced silencing complex
RRBS	Reduced representation bisulfite sequencing
SILAC	Stable isotope labelling by amino acids in cell culture
siRNA	Small interfering RNA
TE	Transposable elements
TIR	Terminal inverted repeats
TSD	Target site duplications
TSS	Transcription start site
ZBED	Zinc finger, BED-domain containing
ZC3H11A	Zinc finger CCCH-type containing 11A

1 Introduction

1.1 Gene regulation

Gene regulation is a fundamental process in viruses, prokaryotes and eukaryotes. It is widely considered that Jacque Monod and Francois Jacob established the basic principle of gene regulation in 1961. Their pioneering work on lac operon of *E.coli* identified that enzymes involved in lactose metabolism are expressed only in the presence of lactose and the absence of glucose. Since then extensive research has been performed to obtain insight into functional importance of gene regulation. It is evident that multicellular organisms contain cells with homogenous genomes but heterogeneous structure and functions due to differential gene expression. Appropriate accomplishment of biological processes including proliferation, differentiation, apoptosis, development and aging is based on proper spatial and temporal gene expression. Now, assembled genome sequences from simple to complex organisms are available to serve as a foundation of comparative genome analysis. This analysis unravelled that the emergence of complex, multicellular organisms is facilitated by dramatic expansion and complexity of gene regulatory network rather than number of protein coding genes. Human genome contains about 20,000-22,000 genes that makes less then 2% of the entire genome. Latest assembly of human genome Hg38 contains 20,364 coding genes (Ensemble, October, 2014). A large part of the genome is involved in gene regulation (Levine & Tjian, 2003).

Gene regulation basically controls the synthesis of gene products including both protein and RNA. Gene regulation is a complex process that is orchestrated by many essential processes mainly including epigenetic regulation, transcriptional and post-transcriptional regulation, as well as translational and post-translational regulation. Epigenetic regulations are essential for development and differentiation. To initiate transcription,

chromatin structure should be open with the DNA template accessible for the transcription machinery. The accessibility is dependent on epigenetic modifications. Epigenetic modifications alter gene expression without changing the DNA sequence and include histone modification and DNA methylation. Histones and DNA collectively forms chromatin and it can be subdivided into euchromatin and heterochromatin. Euchromatin is decompacted and contains transcriptionally active regions of the genome and is located toward interior of the nucleus. Heterochromatin is condensed and transcriptionally silent and is located at the nuclear periphery, near nucleoli or in chromocenters. Chromatin arrangement into open structure is required to carry out transcription. Transcription is a DNA-dependent, RNA polymerase-mediated process that controls when and how much RNA is produced. After transcription, mRNA requires post-transcriptional processing which are modifications that include 5'-capping, splicing and polyadenylation to produce the mature mRNA. Different mRNAs within the same cell have different lifetimes and degrade after a certain time. This limited lifetime enables cell to synthesize protein according to its changing needs. mRNA degradation can be achieved by various mechanisms like small interfering RNA (siRNA), microRNA (miRNA) and mRNAs that contain premature stop codons are degraded through the nonsense-mediated decay system. The processed mRNAs are transported out from the nucleus to the appropriate ribosomes where they are translated into protein that subsequently undergoes post-translational modifications.

1.1.1 Transcriptional regulation

Transcriptional regulation is mediated by combinatorial activity of regulatory elements and an enormous number of factors for recruitment and activity of RNA polymerase (RNA POL) (Figure 1). In eukaryotes three types of RNA polymerases exist; RNA polymerase I which transcribes 28S and 18S rRNA, RNA polymerase II which primarily transcribes mRNA, long non-coding RNA, snRNA and miRNA, and RNA polymerase III that transcribes tRNA, 5S rRNA, snRNA and other small regulatory and structural RNA molecules. The transcription activity depends upon many factors including the different RNA polymerase complex itself with their specific general transcription factors. To achieve regulated transcription the RNA polymerase complex requires regulatory transcription factors that function through binding at *cis*-acting regulatory DNA sequences including core promoters with either TATA boxes and or initiator (Inr) sequences, upstream regulatory promoters and distant enhancers, locus control regions, insulators and silencers. Annotated gene promoters in vertebrates are enriched in CpGs. The regulatory transcription

factors can be grouped into transcriptional activators, coactivators, repressors and corepressors. Eukaryotic RNA PolII transcription initiation complex consists of several multisubunit components including the RNA PolII holoenzyme, general transcription factors, various activators and coactivators or repressors and corepressors depending on regulatory context. In addition a large complexity of chromatin remodelling factors including histone acetylases (HATs) and histone deacetylases (HDACs), histone methylases and demethylases. These different complexes bind with the core promoter to form a stable transcription initiation complex ready to initiate the process of transcription that is further regulated by enhancer, transcription factors and various other regulatory factors (Figure 1a,c). Key signal to initiate transcription involves a specific phosphorylation of the C-terminal tail of RNA polymerase itself. The core promoter is generally 30-40 bp up-/downstream of transcription start site (TSS) and proximal promoter is located within 250-300 bp upstream of the core promoter whereas enhancer is located at approximately 1Mb long distance upstream or downstream from the promoter (Maston *et al*., 2006).

Spatial and temporal transcriptional activation and repression plays a vital role in gene regulation. Sequence-specific DNA binding proteins called transcription factors that function either as activators or repressors mediate transcriptional activation. A prototypic transcription factor is modular and consists of a DNA binding domain, a dimerization domain and a regulatory activation and or repression domain. Transcriptional activators usually tether with core promoter and stimulate transcription by participating in the formation of initiation complex, initiation, elongation and chromatin remodelling (Maston *et al*., 2006). Likewise, transcriptional repression is regulated by repressors that are sequence-specific DNA binding, methylated DNA binding or activator binding proteins. Repressors prevent the transcriptional regulation by inhibiting the binding of RNA polymerase to the promoter or by recruiting chromatin remodelling factors (Gaston & Jayaraman, 2003). Silencer elements bind with repressor to fully repress the transcription. Transcriptional activation is usually linked with histone acetylation while transcriptional repression is associated with histone deacetylation. In addition, a large number of different histone methylations are associated with actively transcribed or repressed promoters, enhancers and inducible promoters.

The encyclopaedia of DNA elements (ENCODE) has generated enormous amount of data to decipher gene regulatory networks in the human genome. Their data obtained from 15 human cell lines demonstrated that human genome is cumulatively comprised of 62.1% and 74.7% of processed and primary transcripts, respectively (Djebali *et al.*, 2012). Chromatin Immunoprecipitation

(ChIP) sequencing data obtained from 119 human transcription factors exhibited that co-association of transcription factor was highly context-specific and regulatory networks of transcription factors and their targets were under strong selection (Gerstein *et al.*, 2012). This data showed that transcription factor binding sites were highly evolutionary conserved and were located in GC-rich and nucleosome-depleted regions (Wang *et al.*, 2012). Moreover, functional characterization of binding sites revealed that functional binding sites were highly conserved and were located in close vicinity of TSS where activating sites were closer to TSS than repressive sites (Whitfield *et al.*, 2012).

Transcription factors

Transcription factors are key cellular components and consist of a repertoire of proteins that regulate transcription by interacting with DNA, accessory proteins and RNA polymerases to activate or repress their action. Transcription factors modulate fidelity of transcription either by directly binding with DNA or indirectly via interaction with other DNA-binding transcription factors. It has been established that human genome contains approximately 1700 to 1900 genes encoding transcription factors where about 1400 transcription factors are manually curated (Vaquerizas *et al.*, 2009). Transcription factors can bind with both promoter and enhancer regulatory regions. Transcription factors bind to degenerate sequence motifs of genomic DNA called transcription factor binding sites (Figure 1b). The sequences of these binding sites are dependent on the actual context of the dimers or multimers of transcription factors that bind the sequence. The predicted binding sites are based on both biochemical and functional data and for a majority of transcription factors where both biochemical and functional data are available a core consensus sequence motif can be established.

Transcription factors play a central role in biology and there are several transcription factor families classified on the basis of function and structure. Transcription factor family categorized on basis of structure includes helix-turn-helix, helix-loop-helix, zinc finger, leucine zipper, homeodomain, metal-binding, paired box, fork head and heat shock factors. For a comprehensive review of different transcription factors see (Latchman, 2008; Pabo & Sauer, 1992). The largest known family of transcription factors is the zinc finger-containing transcription factors. Zinc fingers are important eukaryotic gene regulators that coordinate with zinc ion in the form of finger-like appearance to stabilize the structural conformation. Zinc fingers were first identified in frog to facilitate the function of transcription factor IIIA. Another example of a family of transcription factors is the Paired box (PAX) transcription factors,

which contain a paired domain and a partial or complete homeodomain. PAX transcription factors play key roles in early animal development.

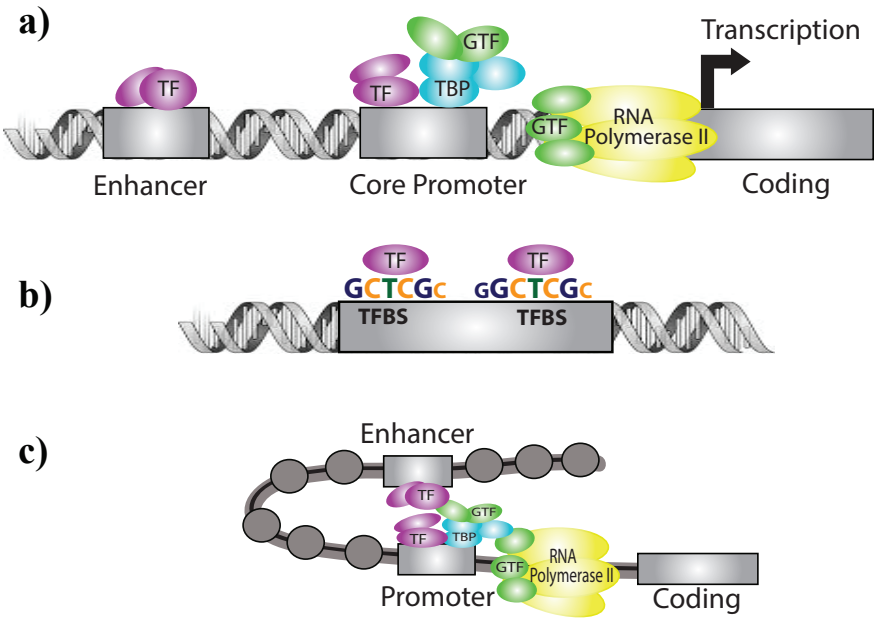


Figure 1. **Transcriptional Regulation.** (a) Transcriptional apparatus encompassing general transcription factors (GTF), and regulatory transcription factors (TF) including activators and repressors, TATA-box binding proteins (TBP) and RNA polymerase II bind with promoter region and start transcription. (b) Transcription factors bind with transcription factor binding sites that are located in genomic DNA. c) A schematic illustration of complex arrangement of eukaryotic transcriptional unit with interactions between TFs binding both promoter and enhancer

ZBED6 is a recently identified novel transcription factor of the zinc finger type that is specific to placental mammals. It is known to act as a repressor of insulin like growth factor II (IGF2) transcription in muscle cells. Earlier studies showed that a single point mutation G to A acted as a causative mutation in a quantitative trait locus (QTL) affecting muscle growth, fat deposition and heart size in pigs (Van Laere *et al*., 2003). Functional analysis showed that this region might be a binding site for an unknown nuclear factor functioning as a transcriptional repressor. Stable isotope labelling of amino acids in cell culture (SILAC) technology was used to identify this unknown and placental mammal conserved nuclear factor. Closer examination showed that the captured peptide was encoded by an intronless gene located in intron 1 of zinc finger CCCH-type containing 11A (ZC3H11A). The open reading frame of this protein

contains more than 900 codons and its protein has two BED domains and one hATC dimerization domain and the sequence showed sequence similarity to DNA transposons. There are already five known members of the ZBED gene family and therefore this newly identified protein was named as ZBED6 (Markljung et al 2009). These findings implied that ZBED6 is a domesticated DNA transposon and belongs to hAT DNA transposon superfamily. ZBED6 binds to consensus sequence of 5′-GCTCGC-3′. ZBED6 is expressed in many tissues such as brain, ovary, muscle, heart and kidney. The widespread expression of ZBED6 suggests its functional importance and that it has the capacity to regulate transcription of both ubiquitously expressed and tissue-specific genes. ChIP-sequencing of mouse myoblasts showed that ZBED6 targets about 1200 annotated genes that are functionally enriched in development, transcriptional regulation, cell differentiation and muscle development (Markljung *et al*., 2009). A recent study revealed ZBED6 involvement in proliferation and survival of pancreatic beta cells. ChIP-sequencing of human islet cells revealed that ZBED6 targeted 351 genes that were associated with transcriptional regulation, apoptosis, biosynthetic processes and transmembrane proteins (Wang *et al*., 2013).

DNA transposons and their domestication in host genomes

DNA transposons are important players in most genomes and they influence several biological processes including the gene regulation process. In 1948, Barbara McClintock discovered the first DNA transposon that was involved in insertion, deletion and translocation of *Zea mays* genome (McClintock, 1953). Later research showed that transposable elements are ubiquitous in the genomes of prokaryotes and eukaryotes. They are important units of genome evolution and are involved in a broad range of cellular functions including gene transduction, gene duplication, exon shuffling, chromosomal rearrangements and new gene creation ultimately leading to genome expansion and evolution (Feschotte & Pritham, 2007). They can be grouped into two major classes on the basis of their mechanism of transposition; class 1 or retrotransposons that transpose via an RNA intermediate and class 2 or DNA transposons that transpose via a DNA intermediate. Eukaryotic DNA transposons can be cut-and-paste transposons like *TC1/mariner, hAT, P-element, PiggyBac* or copy-and-paste transposons such as the *Maverick* and *Helitron* superfamilies. DNA transposons are flanked by terminal inverted repeats (TIRs) and encode at least one enzyme called a transposase that binds with target site duplications (TSDs) and mediates the cut-and-paste transposition reaction. TIRs are followed by TSDs, which are short duplicated

DNA sequences created upon insertion of a transposable element into the host genome.

The origin of DNA transposons remains elusive. DNA transposons make up approximately 3% of the human genome and have shown no activity during the past 50 million years (myr) (Lander *et al.*, 2001). It was estimated that during primate radiation 40 families of DNA transposons were active, however, the primate lineage showed a sudden loss of hAT and other DNA transposon activity around 40 myr ago (Pace & Feschotte, 2007). Later studies indicated a recent activity of various DNA transposable elements in the genome of the brown bat, *Myotis lucifugus* and estimated some of these elements were still expanding in natural populations (Ray *et al.*, 2008).

Transposable elements have been considered as obligate parasites of their host genomes for a long time but recent research has revealed their positive contribution by giving birth to new genes with functions beneficial to the host. This process is referred as molecular domestication (Sinzelle *et al.*, 2009; Volff, 2006; Miller *et al.*, 1997). In contrast to active transposable elements, domesticated elements are single copy and are located at syntenic genomic positions in different species and are devoid of TIRs and TSDs (Sinzelle *et al.*, 2009). The number of genes in the human genome that has originated through domestication of DNA transposons is estimated to be around 50 (Pace & Feschotte, 2007). Based on sequence comparison and phylogenetic analyses *ZBED6* is considered to be domesticated from a *hAT* DNA transposon (Paper 1). The *hAT* DNA transposon superfamily includes both active and domesticated transposons, and is a very large and widespread family in plants, fungi, and animals. The first ever discovered transposon was the Ac transposon in maize by McClintock, which is a *hAT* transposon. The superfamily was named *hAT* after discovery of the related transposable elements hobo in *Drosophilla melanogaster*, *AC* in maize and *Tam3* in *Antirrhinum majus* (Warren *et al.*, 1994). Some other active and well-studied members of *hAT* family are *hermes* from *Musca domestica* and *Tol2* of the Japanese Medaka fish. *hAT* elements contain a conserved domain of 50 amino acids at the C terminal that is essential for transposase activity (Calvi *et al.*, 1991).

1.1.2 DNA methylation

DNA methylation is a crucial epigenetic alteration that regulates many key cellular processes including transcription, chromatin structure, embryonic development, genomic imprinting, X-chromosome inactivation and chromosomal stability. Therefore, proper DNA methylation is essential for mammalian development, and consequently aberrant DNA methylation has been implicated in many pathologies including cancer, imprinting

abnormalities and repeat instability (Robertson et al., 2005). DNA methylation was first discovered in 1948 however its possible role in gene regulation was discovered later. Earlier studies showed that DNA methylation participated in switching on and off the expression of chicken beta globin genes (McGhee & Ginder, 1979). DNA methylation is the main type of DNA modification in mammals and it is a methylation of fifth position of cytosine located in CpG dinucleotide. Additionally, non-CpG methylation is also observed in embryonic stem cells and neuronal cells. Cytosine methylation is a well-studied, a mechanistically understood and a conserved epigenetic modification in most plants, animals fungi and bacteria (Feng *et al.*, 2010; Suzuki & Bird, 2008). This epigenetic modification is deposited and maintained by three main conserved methyltransferases including DNA methyltransferase 1 (DNMT1), DNA methyltransferase 2 (DNMT2), and DNA methyltransferase 3 (DNMT3). These methyltransferases are essential to carry out normal development, and genetic alteration of DNMT1 and DNMT3b leads to 95 % hypomethylation thereby causes loss of e.g. *IGF2* imprinting, silencing of tumor suppressor gene *p16INK4a* and ultimately leads to the growth suppression in colorectal cancer cell line HCT116 (Rhee *et al.*, 2002).

Dazzling advancement in sequencing technologies facilitated to unravel the enormous knowledge about mechanisms and functions of DNA methylation. Human genome is not uniformly methylated and some regions are hypomethylated whereas other regions remain hypermethylated. CpG islands are short interspersed, GC-rich and CpG-rich DNA regions that are frequently located in promoter regions. Most of the genome is hypermethylated at all times and in all cell types. Hypermethylation appears to be the default state and between 60% and 90% of all CpGs are methylated in mammals. Generally, mammalian CpG islands remain hypomethylated at regulatory regions encompassing both promoter regions and distal regulatory regions to remain open and accessible to transcription (Stadler et al., 2011; Straussman et al., 2009) while rest of gene body shows considerably high methylation. DNA methylation epigenetically regulates the gene expression either by repression caused via promoter methylation or by activation mediated by genic methylation (Ball et al., 2009). DNA methylation regulates transcriptional activity through two main mechanisms, firstly by direct inhibition of transcription factor binding, and secondly by recruiting methyl CpG binding protein that may compete with binding of transcription factors. Reduced-representation bisulfite sequencing (RRBS) and DNase I hypersensitive site (DHSs) mapping in a large number of human cells lines and tissues revealed that increased methylation retained inverse correlation with chromatin accessibility and with transcription factor expression by occupying

transcription factor binding sites (Thurman et al., 2012). Hence, these two methods illuminate DNA methylation-mediated transcriptional attenuation. Transcriptional regulation mediated through DNA methylation by occupying transcription factor binding sites is shown in Figure 2.

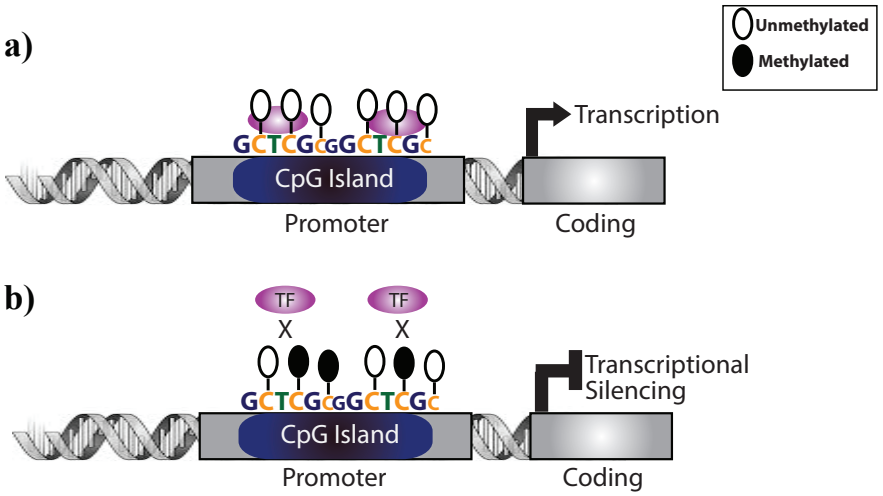


***Figure 2*. DNA methylation-mediated transcriptional silencing.** A schematic illustration of transcriptional attenuation mediated through DNA methylation by occupying transcription factor binding sites located in promoter CpG Island. (a) Unmethylated transcription factor binding sites allow transcription (b) whereas methylated transcription factor binding sites disrupts binding of transcription factors thus leading to transcriptional silencing

Growing interest in epigenetic studies has shed light on the critical role of DNA methylation in the development of cancer. Hypomethylation of oncogenes and hypermethylation of tumor suppressor genes are key features in the progression of cancer. Global hypomethylation of entire genome and gene specific hypomethylation are extensively observed in tumorigenesis and these two phenomena are involved in both benign and malignant tumors. Epigenetic and genetic alterations interplay to predisposition of cancer. However, DNA methylation is a reversible process that indicates its potential as a therapeutic agent to treat a variety of cancer.

Differentially Methylated regions (DMRs) are functional genomic regions that regulate gene transcription and show different methylation profiles among multiple tissues, cells or individuals. For instance, DMRs among different

tissues are called T-DMRS, DMRs between normal and tumor samples are known as C-DMRs.

1.1.3 microRNA

microRNAs (miRNAs) are abundant and conserved small RNA molecules that belong to a large family of small noncoding RNAs. This family encompasses many other known small RNAs including small interfering RNAs (siRNAs), repeat associated small interfering RNAs and piwi- interacting RNAs (piRNAs). miRNA are ~22 nucleotides long and is a well-established class of small RNAs that contributes to regulate gene expression in living organisms. miRNAs mediate gene regulation either through base pairing with 3' UTR of mRNA or by repressing synthesis of protein by poorly understood mechanisms. miRNA biogenesis is started from precursor molecule (pri-miRNAs) that is encoded by independent miRNA genes or from miRNAs located within introns of protein coding genes. Pri-miRNA is further processed by Drosha and Dicer enzymes. This complex cuts pri-miRNA into a shorter sequence of 70 nucleotides to give rise to pre-miRNA that is exported to cytoplasm by exprotin proteins. Dicer cleaves pre-miRNA into ~22 nucleotide to yield a mature miRNA. Mature miRNA coordinates with Argonaute proteins and participates in RNA induced silencing complex (RISC) to mediate transcriptional silencing (Filipowicz *et al.*, 2008) (Figure 3).

Discovery of *lin-4* and *let-7* miRNAs in *C. elegans* opened new horizons to understand the complexity of gene regulation. Later research revealed presence of miRNA in higher organism lineages and simple genomes as well, including plants, animals, viruses and green algae. miRNA family repertoires evolved as a global expansion in mammals and large proportion of recently evolved miRNAs was lost during evolution (Meunier *et al*., 2013). Generally, animal miRNAs interact to target mRNAs with the help of a conserved region named the seed region. Contrary to plants, animal miRNAs are allowed to have imperfect complementarity to recognize their targets. Earlier studies identified that more than half of human genes possess conserved 3'UTR to facilitate pairing with miRNAs and subsequently to control gene regulation (Friedman *et al*., 2009). miRNA attachment with targets leads to deadenylation and consequently degradation of mRNAs (Eulalio *et al*., 2009). However, miRNA can regulate gene expression independent of mRNA degradation. Studies in *Drosophila melanogaster* and *Danio rerio* revealed that miRNAs repressed the translation by disturbing initiation of translation *(Bazzini et al*., 2012; Djuranovic *et al*., 2012). A recent study explained the mechanisms of miRNA mediated gene expression silencing in detail and described it can be mediated through cap inhibition, ribosome 60S joining inhibition, inhibition of

elongation, ribosome drop-off, cotranslational protein degradation, sequestration of P-bodies, mRNA decay, mRNA cleavage and transcriptional inhibition mediated by chromatin reorganization (Morozova *et al*., 2012).

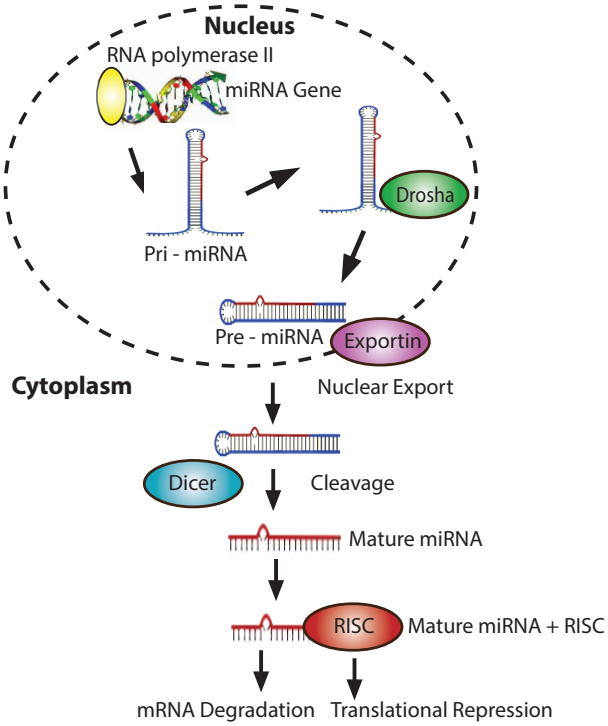


Figure 3. **Biogenesis of miRNA.** miRNA gene is transcribed by RNA polymerase II and produces hairpin structure called pri-miRNA that is further processed by Drosha and Dicer. Exportin proteins help in migration of pre-miRNA from nucleus to cytoplasm where miRNA is cleaved to form mature miRNA that forms a complex with RISC to carry out mRNA degradation and translation repressor

Recent advances in sequencing technologies unveiled widespread expression of miRNAs in different tissues that suggests their participation in a large number of biological processes, including canonical pathways and their activities also influence the development of many diseases as well. For instance, the two first discovered miRNAs, *Lin-4* and *let-7,* are responsible for controlling temporal development in *C. elegans* (Lin *et al.*, 2003; Lee *et al.*, 1993), *miR-14* controls programmed cell death and fat metabolism (Xu *et al.*, 2003), *miR-1* is known to take part in myogenesis and angiogenesis (Sokol & Ambros, 2005; Zhao *et al.*, 2005). Hence, miRNAs have functional roles of

crucial importance in living organisms. Many miRNAs contribute to induction of important cellular signaling like insulin, notch and EGF signaling pathways (Boehm & Slack, 2005; Li & Carthew, 2005; Yoo & Greenwald, 2005). Moreover, role of miRNAs are widespread among all different cell types and are also found to be associated with many disorders including cancer and heart diseases. In short, miRNA-mediated regulatory networks are instrumental in the evolutionary complexity of living organisms.

1.2 Chicken as a growth model

Chicken is a major source of protein for humans and is consumed for eggs and meat production. Domestic Chicken (*Gallus gallus*) is a widely domesticated fowl and belongs to the *Galliformes* order, genus *Gallus* that contains four wild species including red jungle fowl (*Gallus gallus*), grey jungle fowl (*Gallus sonneratii*), Sri Lankan jungle fowl (*Gallus lafayettii*) and green jungle fowl (*Gallus varius*). There are two viewpoints about evolutionary history of domestic chickens; one states they are derived from single wild species of red jungle fowl whereas other believes on polyphyletic ancestry (Kanginakudru *et al.*, 2008; Fumihito *et al.*, 1996). Most available data supports that red jungle fowl is the main ancestral species. However, evidence for a hybrid origin of domestic fowl from grey and red jungle fowl was provided recently (Eriksson *et al.*, 2008). Chickens have a long and proud history being extensively used to explore genetic basis of phenotypic variation. Numerous features encompassing diversity in phenotype, easy and cheap breeding, closest taxonomic group to mammals, high recombination rate and high genetic diversity make chicken a highly suited model organism to study genotype-phenotype relationships (Siegel *et al.*, 2006).

Artificial selection produced chicken breeds with diverse range of morphological, physiological and behavioural phenotypic traits. One example of artificial selection is growth-selected chicken breeds where two growth-selected chicken lines were generated from a founder population of Plymouth Rock chickens. After 50 generations, high weight selection (HWS) lines vary from low weight selection (LWS) lines with a twelve-fold difference in body weight (Pettersson *et al.*, 2013; Marquez *et al.*, 2010; Dunnington & Siegel, 1996) (Figure 4). These two lines exhibit a larger difference in various traits including growth rate, appetite, fat deposition and metabolic traits. HWS chickens are hyperphagic while HWS show low appetite leading to anorexia. QTL mapping of a large intercross population between these two extremely divergent lines identified that there were 13 loci associated with a large phenotypic difference in growth (Jacobsson *et al.*, 2005). Strong epistasis

among these QTLs also contributed to influence the growth (Carlborg *et al*., 2006). Genome-wide single nucleotide polymorphism (SNP) analysis revealed more than 50 regions that have been fixed for alternative alleles in these growth-selected lines (Johansson *et al*., 2010). These outcomes imply that HWS and LWS chickens provide an excellent model to explore complex genetic mechanisms of growth.

Figure 4. **HWS and LWS Chickens.** This figure is clearly illustrating the body size difference between two chicken weight-selection lines (Photo: Paul B. Siegel).

1.3 High throughput approaches

Advances in next generation sequencing (NGS) technology has provided ample amount of high throughput data to study complex regulatory systems in the living organisms. However, data analysis is often very time consuming and requires special bioinformatics expertise to obtain accurate knowledge from high throughput data.

ChIP-sequencing combines chromatin immunoprecipitation with massively parallel NGS of DNA and is used to detect protein-DNA interaction in cells grown in culture or from freshly prepared tissues. It is consequently a technology that is able to detect interactions between transcription factors and target DNA *cis*-elements *in vivo*. ChIP-sequencing can also be used to reveal genomic regions associated with certain histone modifications where antibodies specific for certain histone modifications are utilized. DNA regions containing protein-DNA interaction are enriched with a cross-linker like formaldehyde. DNA strands are sheared by sonication and a protein-specific antibody is applied to capture regions of interest. This complex is pulled down

with beads and crosslinking is reversed and afterwards ChIP DNA is purified. Adaptors are attached with DNA that is bound with protein of interest and subjected to NGS. The large amount of data produced from sequencing requires high computational resources and expertise. Sequencing data is aligned with the reference genome and then peaks and enriched regions are identified with the help of different peak finding tools including MACS , FindPeaks, HPeak and PeakSeq. In order to filter out false positive peaks, reads close to centromeric gaps and repeated regions are removed. Peak finders produce 'wiggle' files that can be imported to genome browsers to directly visualize the peaks. Generally, identified peaks are several hundred bases long and therefore, motif-finding tools like MEME, oPOSSUM, TRANSFAC professional are used to identify optimal binding sites. These motif finder tools search for motifs on the basis of knowledge obtained from transcription factor binding sites databases such as JASPAR and TRANSFAC.

RNA sequencing can address diverse population of RNAs including total RNA that contains all type of RNAs and small RNAs that include miRNA, tRNA and ribosomal profiling. In contrast to microarray technology, RNA sequencing is unbiased and able to detect novel transcripts, sequence variation between alleles, splice variants, alternative transcription start sites and differential expression simultaneously. RNA is isolated and cDNA libraries are prepared that are subjected to NGS sequencing. After sequencing enormous amount of short reads from one end (single end sequencing) or both end (Pair-end sequencing) are obtained. Reads are mapped to reference genome by using Bowtie or BWA tools and transcriptome analysis is performed by applying various tools like TopHat and Trinity. The methods of miRNA- sequencing analysis are similar to RNA-sequencing except for a few differences. miRNA data usually requires single end sequencing and also needs different tools like miRDEEP and miranalyzer for expression analysis. After quantification of RNA expression, differential expression is measured by using tools like cufflinks, DESeq and EdgerR.

Genome-wide DNA methylation analysis is usually performed by bisulfite sequencing and microarray technologies. Microarray provides an inexpensive medium to analysis whole genome methylation of complex organisms. Illumina Infinium DNA methylation 450 assay covers almost all CpG dinucleotide of annotated genes in the human genome. In this technology, genomic DNA is subjected to bisulfite conversion where unmethylated cytosines are converted to uracil. Converted DNA is amplified and applied to chip containing methylated and unmethylated beads where DNA hybridizes with the beads. Later, chips are scanned to measure intensity between two beads. Methylation intensity values help to decide for methylation status.

2 Aims of thesis:

The overall aim of this thesis was to explore the roles of various essential mechanisms involved in regulatory networks of complex traits in vertebrates by adapting computational methods to analyse high throughput NGS data.

The specific aims were to:

I Establish evolutionary relationships among different *ZBED* genes and identify the closest relative of ZBED6. Determine if transposon domestication results from single or multiple domestications.

II Dissect the functional significance of ZBED6 in myoblasts cells. Identify ZBED6 targets other than *IGF2* and the underlying regulatory mechanisms by integrating genome-wide functional analysis.

III Identify potential role of ZBED6 in cancer by scanning DNA methylation data of colon adenocarcinoma. Measure correlation between DNA methylation and RNA sequencing data of ZBED6-targeted regions to identify effects of DNA methylation on transcriptional regulation.

IV Study the role of miRNAs in growth of two extremely divergent chicken weight selection lines. Investigate differential expression of miRNAs between HWS and LWS. Identify DE miRNAs target mRNAs and their underlying biological processes and pathways.

3 Present Investigations

3.1 ZBED evolution: repeated utilization of DNA transposons as regulators of diverse host functions (Paper 1)

3.1.1 Background

Contribution of transcriptional regulation in genomic and phenotypic complexity makes it a key research priority in genome biology. ZBED6 is an important transcription factor and recognized to be originated from a domesticated DNA transposon. In addition to transposons well-established parasitic activities, they are viewed as important drivers in shaping the genomes. Research indicated that transposons play a key role in evolution of eukaryotic gene regulation by acquiring beneficial functions for their host. Transposons participate in transcriptional and post-transcriptional regulatory networks for instance introducing a new transcription start site or transcription factor binding site or miRNA binding site, disrupting existing *cis*-regulatory elements, accumulation of heterochromatin and importantly birth of new transcription factors (Feschotte, 2008). Analysis of 2000 human promoters identified that almost 25% promoters were originated from transposable elements (TEs) and contained many experimentally characterized *cis*-regulatory elements (Jordan *et al.*, 2003). Later research including analyses of promoter regions of all annotated human genes indicated that 83% of the promoters comprised of transposon-derived sequences. 7 Mb of the human genome comprising of 280,000 putative regulatory element has been detected to be exapted from transposon insertions (Lowe & Haussler, 2012). Hence, TEs are main contributors to the wiring of regulatory networks.

ZBED6 is a member of *ZBED* gene family that contains five other known members *ZBED1*, *ZBED2*, *ZBED3*, *ZBED4* and *ZBED5*. All of these genes contain a DNA-binding BED domain with a signature of $Cx_2Cx_nHx_{3-5}$ [H/C] to form a zinc finger. The BED domain was identified by a sequence analysis of plant, animal and fungal proteins and was named BED finger after domesticated *Drosophila melanogaster* BEAF and DREF proteins (Aravind, 2000). *ZBED* genes show widespread expression among vertebrate tissues and perform diverse functions. ZBED1 regulates cell cycle-dependent transcription of various genes encoding ribosomal proteins and is linked to cell proliferation (Yamashita *et al.*, 2007). ZBED3 interacts with axin that activates *wnt*/β catenin signaling, subsequently that regulates both embryogenesis and carcinogenesis in mammals (Chen *et al.*, 2009). A recent study indicated participation of ZBED3 in type 2 diabetes mellitus and detected its *in vivo* regulation by glucose and insulin and its *in vitro* effect on insulin signaling pathway (Jia *et al.*, 2014). ZBED4 is expressed in cone photoreceptors and glial müllar cells of mammalian eye retina, thereby suggesting its potential involvement in retinal morphogenesis and müllar cell functions (Farber *et al.*, 2010; Saghizadeh *et al.*, 2009). ZBED4 is also known to interact with genes with functional roles in hormonal pathways in ovary and several other tissues (Mokhonov *et al.*, 2012). ZBED6 acts as a transcriptional repressor for *IGF2* expression and its role in muscle growth is well studied as discussed in paper 3 (Jiang *et al.*, 2014; Markljung *et al.*, 2009). ZBED6 also contributes to gene expression, proliferation and cell death of pancreatic beta cells (Wang *et al.*, 2013). To date, function of ZBED2 is unknown. Combined, the functional activities of the known ZBED proteins imply that the *ZBED* gene family encodes transcription factors that perform diverse and crucial functions in vertebrates. However, the evolutionary relationship among different *ZBED* genes and *hAT* superfamily is unclear as well as the history of *ZBED* domestication. It is also interesting to identify the closest relative of ZBED6. Therefore, the present study was conducted to investigate these questions related to evolution of *ZBED* gene family.

3.1.2 Results and Discussion

All *ZBED* genes and related DNA transposons were extracted from GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and Repbase (http://www.girinst.org/repbase/) databases by using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and aligned by using clustalx (Thompson *et al.*, 1997). It has been found that BED domains are domesticated from transposons (Aravind, 2000). *ZBED* genes are thought to be exapted from *hAT* DNA transposons, a large superfamily of DNA

transposons that has been recently classified into two families named AC and buster (Arensburger *et al*., 2011). Most domestication events occurred early during evolution and consequently domesticated genes should be located at syntenic genomic positions in various genomes. Therefore, the first analysis was to confirm synteny of all included *ZBED* genes. We found all *ZBED* genes were present at orthologous genomic positions.

ZBEDs phylogeny was reconstructed by using both Bayesian and maximum likelihood methods. For Bayesian inferences MrBayes (Ronquist & Huelsenbeck, 2003) and BEAST (Drummond *et al.*, 2012) were employed and maximum likelihood methods were implemented by using RAXML (Stamatakis, 2006). MrBayes uses Bayesian estimations to reconstruct phylogeny under different stochastic evolutionary models and provides opportunity to analyse heterogeneous data sets including morphological, nucleotide and protein data. BEAST also uses Bayesian estimations for analysing phylogeny but it also infers time-measured phylogenies using strict or relax molecular clock methods. Randomized Axelerated Maximum Likelihood (RAxML) provides substantial substitution models under maximum likelihood methods to infer phylogeny for larger data sets. DNA and protein phylogenetic trees obtained from Bayesian inferences and maximum likelihood methods were highly similar. Our analysis reveals that all ZBED*s* belong to *AC* family except *ZBED5* that is a member of Buster family. Since the BED domain is present in both *AC* and *Buster* family members we therefore suggest that all domesticated transposons belonging to the *AC* and *Buster* family should be named as *ZBED* genes. *ZBEDs* showed two monophyletic clades in *AC* transposon family. *ZBED1* gene from all species form one clade whereas *ZBED6*, *ZBED2*, *ZBED3*, *ZBED4* are part of another monophyletic clade. We identify a new member of the *ZBED* family that is called *C7ORF29* sharing sequence similarity with frog *ZBEDx* and this gene is phylogenetically the most closely related to *ZBED6*. Active transposons of *AC* family are ancestors of these two monophyletic clades. We have two main hypotheses for transposon domestications that gave birth to *ZBED* gene family. First hypothesis states that two separate domestication events followed in one case by sequential duplication created *ZBED* genes and second hypothesis describes that single domestication event followed by successive rounds of duplications gave rise to this transcription factor family (Figure 5). Our data favours hypothesis one. Presence of diverse invertebrate taxa between *ZBED* clades suggested that there were at least two independent transposon domestication events that occurred in primitive jawed vertebrate ancestors about 400 mya. These findings are consistent with molecular clock estimates for coalescent dates estimated in BEAST (Drummond *et al.*, 2012).

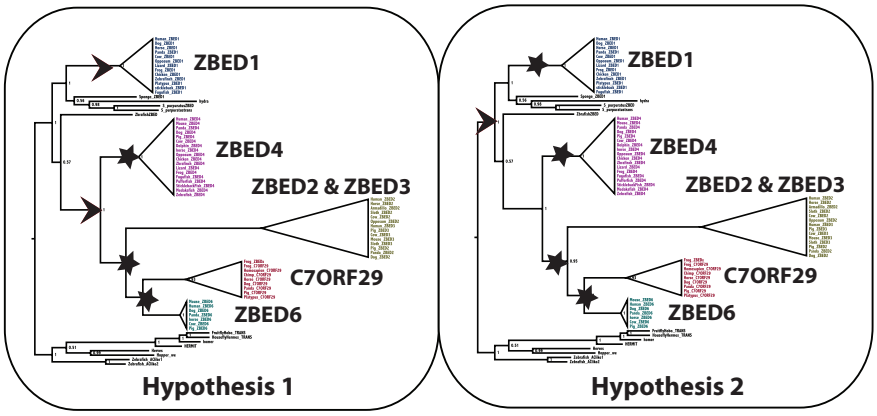


Figure 5. **Hypotheses of transposon domestication.** A schematic illustration of two hypotheses for transposon domestication. Hypothesis one demonstrates that at least two separate domestication events followed in one case by sequential duplications were responsible for the birth of *ZBED* gene family. Hypothesis two states that a single domestication event followed by successive duplications led to arise of *ZBED* gene family.

ZBED genes encode single or multiple DNA binding BED domains and a catalytic domain that further harbours alpha helical and *hAT* dimerization domains. The catalytic domain possesses residues that are critical for transposition process of transposons. *C7ORF29*, the newly identified member of this family, shares sequence identity with *ZBEDX* that is currently known to be present in *Xenopus tropicalis* only. *C7ORF27* encodes a single truncated protein with catalytic domain and the locus shares genomic synteny with *ZBEDx* in human genome. ZBED1, ZBED2 and ZBED3 contain a single BED domain, whereas ZBED6 and ZBEDX contain two BED domains and ZBED4 has four BED domains. C7ORF29 has lost its BED domain during the course of evolution. We also investigated evolutionary events that caused variation in the number of BED domains among the *ZBED* genes. Our analysis suggested that multiple duplication events occurred after *ZBED* gene domestication. The BED domains encoded by the same *ZBED* gene are most closely similar to each other compared with BED domains of other *ZBED* genes. This supports independent duplication events after domestication rather than recombination events. Transposon domestication process may undergo adaptive evolution followed by stabilizing selection as observed in mammalian centromere associated proteins-B (Casola *et al.*, 2008). We also measured the presence of signatures of selection in our data by using PAML (Yang, 2007). This tool uses different methods to estimate synonymous and non-synonymous rates between protein-coding DNA sequences, to infer positive selection and also to perform

reconstruction of ancestral sequences. We could not find evidence of positive selection acting on these *ZBED* genes. Long branches, high level of divergence and saturation of synonymous mutations made it difficult to measure adaptive evolution for our data using existing models.

Transposons are important drivers of complexity in genome evolution. Our findings highlight the propensity of transposons to give rise to new host proteins of diverse regulatory functions among vertebrates. It is tempting to speculate that the DNA binding activity of transposon-encoded proteins have been an important feature during the selection for sequences used to increase the number of diverse DNA-binding transcription factors, recombinases and integrases as well as other proteins whose functions are dependent on DNA interactions.

3.1.3 Future perspectives

The predicted structure of ZBED proteins showed strong similarity with the crystal structure of hermes, a housefly transposase (Hickman *et al.*, 2005). ZBEDs have conserved DDE motif and W motif (Figure 4a). The crystal structure of Hermes, revealed that D180, D248, E572 triad is located in catalytic domain and W319 resides in inserted domain. These conserved acidic amino acids organize in a suitable conformation to facilitate the catalytic activity of transposition process (Hickman *et al.*, 2005). DDE catalytic triad conservation in ZBEDs is also consistent with RAG recombinases that belong to another important gene family domesticated from DNA transposons and the retroviral integrase superfamily. Mutation in these acidic residues leads to halt or reduce the catalytic activity (Zhou *et al.*, 2004). ZBEDs retain conservation in CxxH motif that is analogous to hermes transposase and other active transposases. This motif is located after second D of catalytic triad and is important for transposase catalytic activity (Zhou et al., 2004). The functional importance of these critical residues in ZBED proteins remains to be elucidated. Besides these known motifs we identified a highly conserved motif consisting of a three amino acids motif, LDP of unknown function (Figure 4b). This motif resides after W319 of hermes. Site-directed mutagenesis may help to investigate the functional importance of these conserved residues in ZBEDs.

Another motif that is observed in ZBED3 sequences is the seven residues long motif PPPPSPT, which is known to serve as an axin binding site (Chen *et al.*, 2009) The serine and threonine residues of this motif are believed to be phosphorylated to enhance the interaction with axin. Our data showed that it was present only in mouse ZBED3 sequence while other orthologues (human, cow, pig and sloth) showed presence of a partial motif (PPPP), and the

potential residues for phosphorylation (serine and threonine) were missing (Figure 4c). The underlying function of ZBED3 carrying this partial motif (PPPP) can only be understood by performing functional analysis.

As discussed above, ZBED6 is involved in gene regulation of pancreatic beta cells and muscle cells and ZBED3 is known to regulate insulin-signalling pathway, both of these *ZBED* genes are also close relatives. This implies these two *ZBED* genes may work together in similar pathways to regulate gene expression in pancreatic beta cells. Experimental verifications will help to unveil their potential cooperative activity.

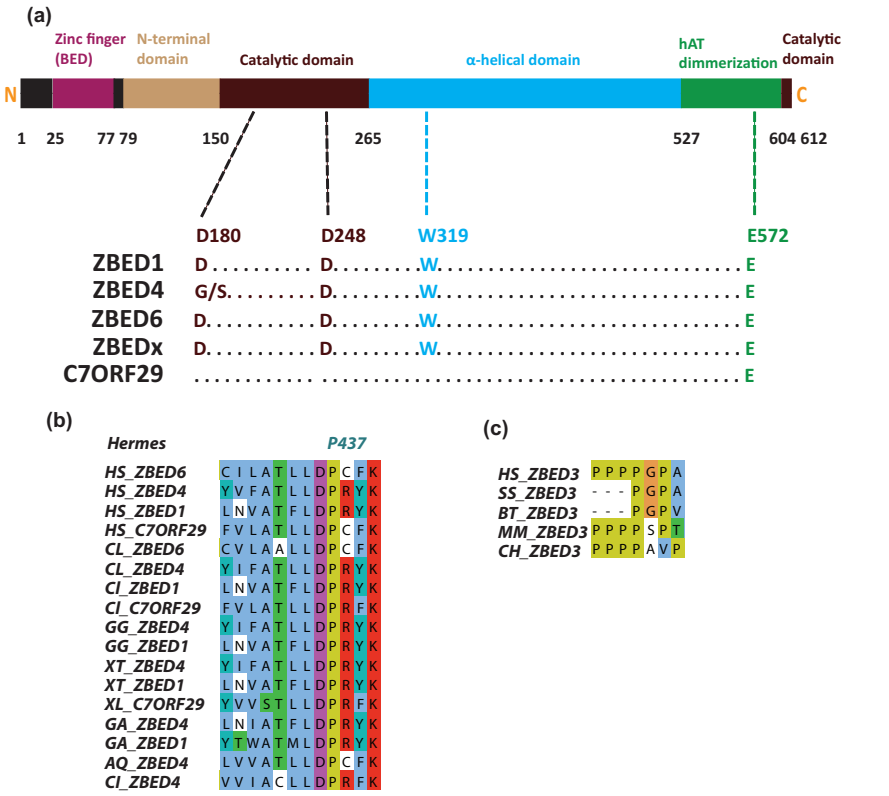


Figure 6. (a) **Comparative annotation of ZBED sequences and C7ORF29 with molecular architecture of hermes transposase.** Hermes transposase molecular architecture is represented here, DDE (catalytic triad) and W (active site) conservation in ZBED proteins is highlighted **(b) Conserved motif in ZBED sequences. (c) PPPP(S/T)PX(T/S) motif in ZBED3**

AQ = *Amphimedon queenslandica*, BT = *Bos taurus*, CH = *Choloepus hoffmanni*, CI = *Ciona intestinalis*, CL = *Canis lupus*, GA = *Gasterosteus aculeatus*, GG = *Gallus gallus*, HS = *Homo sapiens*, MM = *Mus musculus*, SS = *Sus scrofa*, XL = *Xenopus laevis*, XT = *Xenopus tropicalis*

3.2 Genome-wide analysis of the DNA methylation landscape in colon adenocarcinoma reveals hyper-methylation at ZBED6-target regions (Paper 2)

3.2.1 Background

As described above, ZBED6 is a novel transcription factor that acts as a transcriptional repressor of *IGF2* mRNA expression in skeletal muscle cells. Bisulfite sequencing of 300bp region centered around the Quantitative Trait Nucleotide (QTN) in pigs indicated DNA hypermethylation of this CpG island harbouring a ZBED6 binding site in liver. Further, electrophoretic mobility shift analysis unravelled that methylation of this region hampered ZBED6 affinity with its binding site (Van Laere *et al.*, 2003). ChIP-sequencing data of ZBED6 in myoblast C2C12 cells demonstrated functional enrichment of various disorders particularly cancer (Markljung *et al.*, 2009). A recent research conducted by our group revealed ZBED6 binding sites showed a strong association with active promoter marks including H3K4me2, H3K4me3 and H3K27ac histone modifications (Jiang *et al.*, 2014) (Paper III). Hence these findings illustrate participation of ZBED6 target genes in cancer and their association with epigenetic alterations *i.e.* correlation with histone modifications and DNA methylation-dependent ZBED6 binding.

In previous ChIP-sequencing studies we have shown that ZBED6 interacts with several genes with known functions in cancer. Therefore, we decided to further analyse DNA methylation patterns in colorectal cancer as a model. Cancer is a complex disease that is caused by both genetic and epigenetic alterations. Role of DNA methylation in predisposition of cancer is well established. Almost all types of cancers are affected by aberrant DNA methylation. Colorectal cancer is the third most prevalent cancer type and the relevance of differential DNA methylation in the development of this cancer has been extensively studied. Studies unveil that hypermethylation of CpG islands located in promoter regions is closely associated with transcriptional silencing and consequently it contributes to altering critical signal transduction pathways in colorectal cancer (Baylin & Ohm, 2006; Toyota *et al.*, 1999). Besides the well-known role of DNA methylation of promoters and CpG islands, differential DNA methylation of CpG island shores has also been observed in different tumors compared to the adjacent normal healthy tissue. A CpG island shore is defined as a sequence located within a 2kb distance from a CpG island (Irizarry *et al.*, 2009). Tissue-specific differential DNA methylation was observed to be more extensive in CpG island shores compared to CpG islands and it was highly correlated to change in gene expression.

These conserved tissue-specific CpG island shores showed similar percentage of hypomethylation and hypermethylation in colon cancer. These regions, displaying differential methylation in cancer, were named cancer differential methylated regions (cDMR) (Irizarry *et al.*, 2009). Further, cDMRs showed increased stochastic methylation variation in colon, lung, breast, thyroid and Wilms' tumors. Whole genome bisulfite sequencing of colon cancer showed loss of methylation stability at CpG islands, shores and large block of hypomethylation encompassing more than half of the genome (Hansen *et al.*, 2011). Based on these studies it was proposed that modifications in DNA methylation play an important role in progression of colon cancer.

Transcription factors show variable binding efficiency depending on whether or not their DNA binding sites are methylated. Several transcription factors including AP1, CREB, E2F and NFKB showed disruption of binding after DNA methylation of their binding sites (Singal & Ginder, 1999). In contrast, other transcription factors such as KLF4, HOXA5 and GATA4 showed capacity to bind with methylated CpGs (Hu *et al.*, 2013). General transcription factors including SP1, NFY and YY1 were capable of cooperative binding and resistance to *de novo* methylation of CpG islands in colorectal cancer (Gebhard *et al.*, 2010). Recently, the ENCODE research project which was carried out in a large number of human cell lines and tissues encompassing cancer cell lines as well, demonstrated a negative relationship between DNA methylation and transcription factor binding sites (Thurman *et al.*, 2012). Hence, these studies shed light on the relationship between DNA methylation and transcription factor binding ability.

In short, DNA methylation has a widespread role in the development of colon cancer and the DNA-binding capacity and functionality of transcription factors are correlated with DNA methylation. As discussed above, ZBED6 binding sites contained functional enrichment for cancer-associated genes and ZBED6 showed DNA-methylation dependent binding, based on this knowledge we hypothesized a potential role of ZBED6 in colon cancer. Colon cancer development is regulated by CpG island methylator phenotype (CIMP) (Colussi *et al.*, 2013) and IGF2 acts as a diagnostic marker for CIMP due to its promoter hypermethylation (Weisenberger *et al.*, 2006), which provide additional motivations to perform this study. To obtain further knowledge regarding the complex relationship between ZBED6 binding and DNA methylation and ultimately its implication in cancer, we screened publically available genome-wide DNA methylation data for colorectal cancer.

3.2.2 Results and Discussion

Earlier studies showed that the *IGF2* region around the ZBED6 binding site in intron 3 was methylated in liver thus disrupting binding of ZBED6 in this tissue (Van Laere *et al.*, 2003). Therefore, we started our analysis through investigating methylation patterns of 300 bp region centered around ZBED6 binding site in *IGF2*. We observed DNA methylation of this region for all available TCGA tumor data that has more than 10 normal tissues. This 300 bp region in intron 3 of *IGF2* contains three probes of Illumina Infinium Human DNA methylation 450 that includes methylation states (β values) of 480K CpG sites. DNA methylation of these probes in normal and tumor tissues exhibited remarkable methylation difference in Colon adenocarcinoma (COAD), Lung adenocarcinoma (LUAD), Rectum adenocarcinoma (READ) and Stomach adenocarcinoma (STAD) (Figure 5). Later, we observed the methylation pattern of all probes residing in *IGF2*, and this analysis showed that COAD had a strikingly different pattern of methylation around QTN region compared with the rest of *IGF2* probes in tumor versus normal tissues (data not shown). Therefore, we have selected COAD to observe the DNA methylation profile of ZBED6 binding sites. COAD DNA methylation data includes 420 tumor and 75 normal tissues. Previously defined 2469 *bona fide* ZBED6 binding sites in human genome were analysed in this study.

Two data sets, 500bp regions that are +/-250bp on each side of the ZBED6 core binding site and 5 bp that represent the core binding site were created from this data. Wilcoxon rank-sum test on 500 bp ZBED6 binding region containing 8179 probes and on control region of randomly selected 8150 probes was performed. This analysis showed that there were a total of 1170 ZBED6 targeted probes with significantly different methylation pattern in COAD tumor versus normal tissues. In order to observe differential methylation Δβ was calculated for putative ZBED6 binding regions and control data. ZBED6 binding region probes were significantly hypermethylated compared with the control probes. After taking into consideration the functional genomic distribution and CpG content, the difference in number of hypermethylated and hypomethylated probes residing in ZBED6-targeted and control probes was not significant. This analysis unravelled that ZBED6-targeted regions were not enriched in differential methylation compared to overall altered methylation pattern in COAD. Furthermore, we evaluated the methylation of ZBED6 core binding sites and revealed that ZBED6 binding sites were hypermethylated in eight genes including *C8orf42*, *TMEM163*, *ZNF347*, *GRID2*, *HFM1*, *ZNF354C*, *NAALAD2* and *CDH6*.

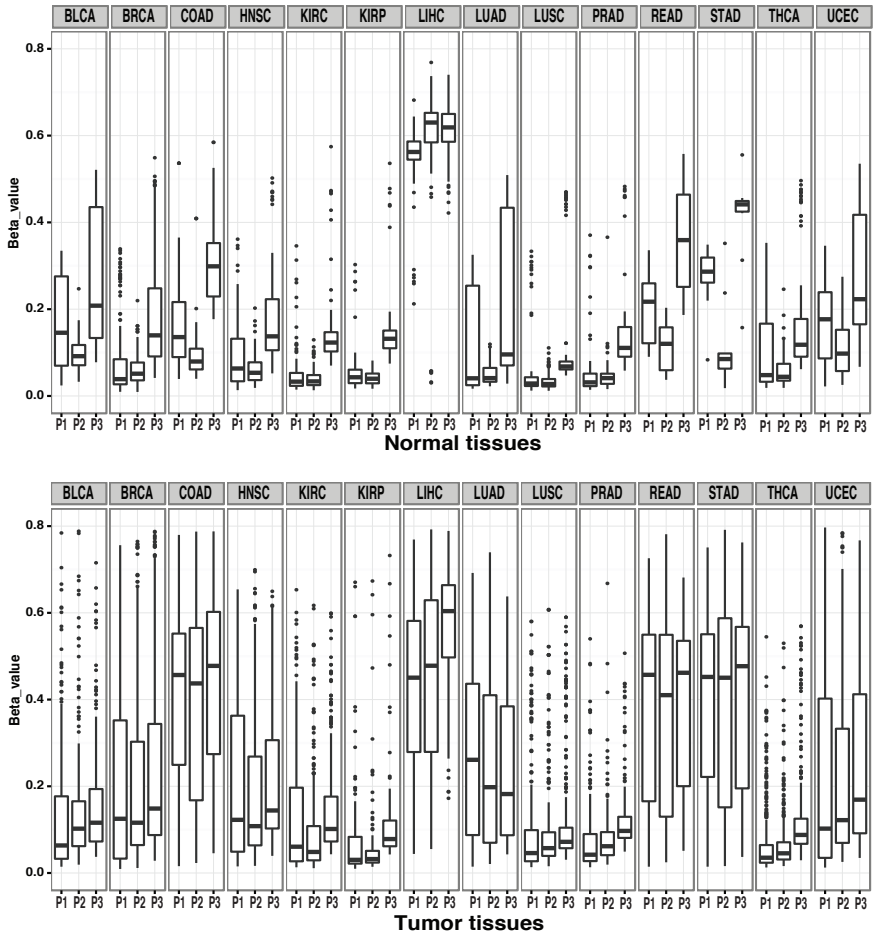


Figure 7. **Methylation patterns around QTN in various tumor and normal tissues.** DNA methylation patterns for the 300 bp region around ZBED6 binding site at the IGF2 gene in normal and tumor tissues of various tumors available at TCGA are shown in these box plots. These tumors are Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Head and Neck squamous cell carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Pancreatic adenocarcinoma (PAAD), Prostate adenocarcinoma (PRAD), Rectum adenocarcinoma (READ), Stomach adenocarcinoma (STAD), Thyroid carcinoma (THCA), and Uterine Corpus Endometrioid Carcinoma (UCEC).

Next, we evaluated Illumina HiSeq RNAseqV2 data for COAD available at TCGA. Here we focused on matched samples containing both methylation and expression data, which was available for a total of 11 COAD samples. Differential expression between tumor versus normal samples was measured for significantly methylated ZBED6-targeted probes in DESeq2. This analysis exhibited that 246 genes are differentially expressed. Pearson correlation revealed that there is a moderate negative correlation between differential methylation and differential expression. There are a total of 46 genes at threshold of $|\Delta\beta| > 0.20$ and that have differential expression. Most of these 46 genes were hypermethylated and showed enrichment in CpG island and promoter regions. Among these 46 genes *NAALAD2* and *ZNF347* showed hypermethylation of core ZBED6 binding sites. Further, we analysed 500 bp region around ZBED6 binding site of these 46 differentially methylated and differentially expressed genes in MEME tool to identify additional binding sites of other transcription factors. Our analysis revealed the presence of putative binding sites for PAX5 in close proximity of ZBED6. Majority of genes showed that these sites are located close to TSS and about 40% of these sites are evolutionary conserved. There are three promoter models for orientation of ZBED6 and PAX5 binding sites with respect to TSS. Two models illustrate ZBED6 and PAX5 binding sites are positioned in same orientation either upstream or downstream of TSS and one model shows that Pax5 is located upstream and ZBED6 is located downstream of TSS. 82% of these promoters of 46 genes exhibited that ZBED6 and PAX5 sites are located in same orientation with respect to TSS. Gene ontology enrichment analysis performed in g:Profiler revealed enrichment of pathways in cancer and PI3-AKT signalling pathway. Our study could not provide clues for enrichment of hypermethylation of ZBED6 binding sites in comparison to overall aberrant methylation profile in COAD. However, our results suggest a role for ZBED6 in COAD by revealing hypermethylation of ZBED6-targeted genes and its correlation with transcriptional attenuation.

3.2.3 Future perspectives

Our results revealed 46 candidate genes exhibiting both differential methylation and differential expression and these genes harboured putative ZBED6 target sites identified by ChIP-sequencing. Functional studies are required to obtain evidence implicating whether or not any of these ZBED6 binding sites are biologically important. Majority of genes showed hypermethylation within 250 bp of ZBED6 sites, hence further assessment of the specific methylation status of CpG sites within ZBED6 core binding site is required as well. Transcription factor binding sites analysis unveiled the

presence of PAX5 binding site in neighbouring region of ZBED6 core binding site. Functional experiments, for instance transient transfection experiments followed by Luciferase assays using promoter constructs containing Pax5 and ZBED6 binding sites in the three different contexts relative to TSS is required to establish their functionality. Moreover, EMSA experiments to define potential cooperative binding between Pax5 and ZBED6 is required to validate this cooperative activity of ZBED6 and PAX5 transcription factors.

Our analysis identified a number of candidate genes containing binding sites for ZBED6 with potential roles in development and progression of COAD that encourages further screening of colorectal cancer to discern a direct role of ZBED6 in development of COAD and whether or not it has a direct role in DNA methylation, to investigate this phenomenon, we have generated DNA methylation data from RKO wild type and RKO ZBED6 knock out cell lines by using Illumina Infinium Human DNA methylation 450 platform. Analysis of these data is on going and will reveal insights concerning ZBED6's potential role in DNA methylation.

3.3 ZBED6 Modulates the Transcription of Myogenic Genes in Mouse Myoblast Cells (Paper 3)

3.3.1 Background

Two independent intercrosses between the European wild boar and the domestic Large White and between Pietrain and Large White pigs were generated for mapping QTLs affecting muscle growth, fat deposition and heart size that resulted in increased muscle mass, enlarged heart size and reduced back-fat thickness. This QTL showed paternal expression and was mapped to close proximity of the paternally expressed *IGF2* gene, therefore this region became positional candidate (Jeon *et al.*, 1999; Nezer *et al.*, 1999). Re-sequencing data from different breeds suggested the presence of a causative mutation that was a single G to A transition at *IGF2*-intron3-3072 and was named QTN (Van Laere *et al.*, 2003). QTN is located in an evolutionary conserved CpG island and 94bp DNA flanking around QTN is evolutionary conserved in placental mammals. Pigs that inherited the mutation showed a 3-fold up-regulation in *IGF2* mRNA expression. Electrophoretic mobility shift assay (EMSA) showed that the mutant allele disrupted affinity with an unknown nuclear factor and the functionality of this differential binding was also confirmed by luciferase assay. Therefore, it was suggested that this region contained a binding site of an unknown nuclear factor, its mutation disrupted the interaction and led to 3-fold upregulation of *IGF2* mRNA expression in skeletal muscle (Van Laere *et al.*, 2003). The mutation in the binding site of this unknown nuclear factor also effects expression of *IGF2* antisense noncoding transcript (Braunschweig *et al.*, 2004). Later studies showed that this was a binding site of a novel transcription factor named ZBED6 (Markljung *et al.*, 2009). Silencing of ZBED6 in mouse C2C12 myoblast effects IGF2 expression, cell proliferation, wound healing and myotube formation.

To define the complexity of binding of ZBED6 in the genome, ChIP-sequencing using anti-ZBED6 antibody was performed in C2C12 cells. This analysis demonstrates that ZBED6 potentially targets 1200 annotated genes of which 262 genes are encoding transcription factors. Majority of these binding sites are located in vicinity of TSS and within or near CpG islands and hold enrichment for a consensus motif 5′-GCTCGC-3′. Gene ontology analysis identified enrichment of ZBED6 targets in various biological processes including development, regulation of transcription, cell differentiation,

neurogenesis and muscle development. ZBED6 targets also showed association with several disorders such as developmental disorders, cancer, cardiovascular disease and neurological disease (Markljung *et al.*, 2009). These results suggested a widespread role of ZBED6 in crucial gene regulatory networks and opened many questions to explore. First, functional studies revealed that ZBED6 acted as a repressor of *IGF2* transcription and ChIP-sequencing data in mouse myoblast exhibited that ZBED6 targeted 1200 annotated genes. However, little is known about the functional role of ZBED6 in the control of transcription of these putative targets. Therefore, it is interesting to elucidate the underlying biological mechanisms and also investigate the potential role of ZBED6 as an activator or as a repressor. Second, transcriptional repression is associated with histone modifications and DNA methylation, so it is important to know what are the mechanisms underlying transcriptional repressive activity of ZBED6. For detailed characterization of the functional significance of ZBED6 and to investigate the first question we generated RNA-sequencing data, and microarray data in ZBED6-silenced and control myoblast cells. ChIP-sequencing analysis with six different histone modification was performed to explore the second question.

3.3.2 Results and Discussion

To elucidate regulatory role of ZBED6 we generated SOLiD strand-specific RNA sequencing data and Illumina BeadChip arrays in triplicate ZBED6-silenced C2C12 cells at two different time points, day 2 and day 4. RNA sequencing data showed that 1,094 and 4,412 genes were differentially expressed (DE) at day 2 and day 4, respectively and 780 DE-genes among them were significantly correlated between two time points. 368 DE-genes were highly correlated with array data. Among these DE-genes there are 20 small nucleolar RNA (snoRNA) genes that show an up-regulation after silencing and the result is confirmed with qPCR. Five out of twenty of these snoRNA genes also harbour ZBED6 binding sites suggesting that ZBED6 is directly or indirectly involved in repression of snoRNA expression. Functional enrichment analysis of 780 DE genes in the Database for Annotation, Visualization and Integrated Discovery (DAVID) identified that these genes are significantly associated with muscle proteins, contractile fibre, myofibril, muscle contraction actin cytoskeleton, actin binding and heart development.

We generated additional ChIP-sequencing data using Illumina HiSeq system to increase resolution and enrichment of existing ChIP-sequencing experiment (Markljung *et al.*, 2009). This data identified 3,780 ZBED6 peaks and enriched in DE-genes in particular up-regulated genes that provided

additional support for a role of ZBED6 as a transcriptional repressor. However, most of these DE genes were not direct targets of ZBED6 and therefore we analysed promoters of DE genes to identify transcription factor binding sites in 2 Kb region flanking TSS in oPOSSUM tool. This tool uses phylogenetic footprinting to scan promoter regions for identification of overrepresented transcription factor binding sites from the JASPAR database on the basis of statistical significance (Ho Sui *et al.*, 2007; Ho Sui *et al.*, 2005). Binding sites for NFKb1, ELK4 and SP1 were significantly overrepresented in down-regulated genes following ZBED6-silencing while MEF2a and PRRX2 were found in promoters of genes that were up-regulated following ZBED6-silencing. qPCR confirmed two-fold decrease in expression of *NFKb1* and *ELK4* mRNA. Hence, these findings imply that ZBED6 contains many functional targets sites in addition to *IGF2*.

22% peaks contained GCTCG core binding site of ZBED6 within 100 bp of the summit and 45 DE genes showed evolutionary conserved ZBED6 binding site where 17 genes were down-regulated and 28 were up-regulated after ZBED6-silencing. Genes which were up-regulated following ZBED6-silencing contained a palindrome sequence followed by the core ZBED6 binding site 5'-GCCTAGGCTCG-3' which was also observed in *IGF2*. In contrast, none of the genes that were down-regulated following ZBED6-silencing showed the palindrome sequence. Among the genes which were down-regulated after ZBED6-silencing we selected *Twist2* for further analysis. To explore the importance of the palindrome and ZBED6 function as an activator versus repressor we performed Luciferase assays using four reporter constructs of the *Twist2* promoter region that contained ZBED6 binding site, a deletion (DEL) construct where the ZBED6 site was deleted, a G to A mutation (MUT), like the mutation in the Igf2 gene and an artificial construct including the palindrome sequence as well. Wild-type *Twist2* reporter was down-regulated after ZBED6-silencing whereas the DEL and MUT constructs were not. The artificial reporter with the palindrome was up-regulated following ZBED6-silencing. These Luciferase experiments as well as EMSA further confirmed that ZBED6 acted as an activator of *Twist2* transcription and showed ZBED6 acted as a repressor in the presence of a palindrome. Twist2 acts as an antagonist of myogenesis by repressing transcription of *MYOD1* and *MEFC2*. Our data suggests ZBED6 may regulate myogenesis by directly repressing *IGF2* and activating the transcription of *Twist2*.

As discussed above, DNA methylation disrupts binding of ZBED6 and therefore to investigate its interaction with histone modification we overlapped ChIP-sequencing data of C2C12 cells for six histone modifications H3K4me1, H3K4me2, H3K4me3, H3K27me3, H3K27ac and H3K36me3 with ZBED6

binding sites. This analysis demonstrated that ZBED6 binding sites were strongly associated with active promoter histone marks including H3K4me2, H3K4me3, and H3K27ac while repressive histone marks like H3K27me3 did not show the enrichment. Our results suggest ZBED6 can bind with histone containing DNA to modulate transcription of actively transcribed genes. Further, ZBED6 binding sites of down-regulated genes exhibited stronger enrichment for H3K4me2 than up-regulated genes. To investigate the role of ZBED6-silencing in histone modification we generated ChIP-sequencing data by using H3K4me3, H3K9Ac and H3k4me2 antibodies in ZBED6-silenced and control C2C12 myoblasts. The data showed overall similar enrichment for all modifications in control and ZBED6-silenced cells. ZBED6-silencing did not show any major impact on H3K4me3, H3K9Ac and H3k4me2 modifications. However, this data confirmed the enrichment of H3K4me3, H3K9Ac and H3k4me2 modifications in up-regulated genes compared to down-regulated genes that reinforced association of ZBED6 with active histone marks. Moreover, qPCR and western blotting in differentiated and undifferentiated C2C12 cells showed that ZBED6 primarily acted as a modulator of expression of *IGF2* and myogenin rather than a classical repressor or activator.

3.3.3 Future perspectives

More ChIP-sequencing and RNA sequencing data obtained from different cell lines and tissues of diverse placental mammals, mainly human and mouse will assist to decipher a widespread regulatory role of ZBED6. This data will help to identify crucial ZBED6 targets that are either activated or repressed by ZBED6. Likewise, functional studies are required to find functional ZBED6 sites and it will facilitate to understand ZBED6 functional implications and underlying complex regulatory networks. It will be extremely useful to identify proteins that interact with ZBED6, this protein-protein interaction can be investigated through yeast two-hybrid screening or by co-immunoprecipitation using anti-ZBED6 antibodies. Additionally, high throughput data including ChIP-sequencing, RNA sequencing and miRNA sequencing obtained from ZBED6 knockout mice will provide an extra layer of critical functional implications related to ZBED6.

3.4 Differentially expressed microRNAs in relation to growth of two extremely divergent chicken lines (Paper 4)

3.4.1 Background

Two extremely divergent chicken lines HWS and LWS represent an excellent model to study the underlying genetic mechanisms responsible for differential growth. To obtain insights into complex phenomenon of growth, many studies have been performed using this model. One study that focused on the hypothalamus of these growth lines indicated that electrolytic lesion in ventromedial hypothalamus influenced feed intake in LWS and they exhibited obesity syndrome whereas HWS showed no impact in feed consumption and body weight (Burkhart *et al.*, 1983). A recent study was conducted to reveal differential gene expression in hypothalamus of growth-selected chicken lines. This genome-wide expression analysis identified differential expression of 585 genes. 97 of these DE genes overlapped with previously identified growth QTLs and selective sweep regions with a significant enrichment. Differentially expressed genes demonstrated functional enrichment in cell cycle, cell death, neuronal plasticity and MAPK signalling and these findings supported a role for DE genes in differential growth (Ka *et al.*, 2011). These outcomes implicate as expected a crucial involvement of hypothalamus in the growth of HWS and LWS birds.

miRNAs function as indispensible modulators of a wide array of biological processes and contribute to organismal development and growth. miRNAs functional roles have been established in chicken embryogenesis (Hicks *et al.*, 2008), adipogenesis (Yao *et al.*, 2011) and myogenesis (Andreote *et al.*, 2014). However, potential role of miRNAs in growth of chicken is still poorly understood and the studies that have been performed focused only on few tissues. miRNA analysis conducted in skeletal muscle of layer and broiler chicken illustrated expression of 33 novel and 189 known chicken miRNAs in this tissue (Li *et al.*, 2011). 17 of these miRNAs were confirmed for differential expression between these two chickens by functional analysis. DE miRNAs target many mRNAs that are involved in myogenesis (Li *et al.*, 2011). Moreover, miRNA expression in skeletal muscle of dwarf and normal chicken was also observed. This microarray analysis showed that *miR-1623*, *miR-181b*, *let-7b* and *miR-128* were differentially expressed in these two types of chicken. Furthermore, it was shown that only *let-7b* was able to bind with 3'UTR of growth hormone receptor mRNA to regulate fat deposition and skeletal muscle growth thereby involved in determining chicken dwarf phenotypes (Lin *et al.*, 2012). These studies stress the importance of miRNA in growth of chicken skeletal muscle. Therefore, we investigated the expression pattern of miRNA

in the HWS and LWS chicken lines to gain increased knowledge about chicken miRNA in general and in particular to investigate for potential differential expression of these miRNAs between tissues and between lines. This model should represent a more powerful RNA-seq experiment involving a comparison from three tissues including hypothalamus, pectoralis major muscle and liver which may shed light on the role of miRNA in chicken growth.

3.4.2 Results and Discussion

Hypothalamus is the master regulatory center for growth control. Liver is the supply organ where IGFs are expressed and exported via the bloodstream to other tissues of the body to regulate cell growth and differentiation. Skeletal muscle is a major demand organ that requires growth hormones to regulate muscle growth and development. To investigate role of miRNAs in this whole growth cycle we generated more powerful data obtained from these three tissues of eight male birds from HWS and LWS growth lines. In total, 24 miRNA samples were subjected to HiSeq2000 (Illumina) sequencing that resulted in reads of 50bp length. Afterwards, reads from 24 samples obtained from hypothalamus, pectoralis major muscle and liver of HWS and LWS were analysed in miRDeep2 to identify novel and known miRNAs expressed in these tissues. The unique feature of this tool is identification of both novel and known miRNAs. miRDeep2 uses Bayesian statistics algorithm to predict canonical and non-canonical miRNAs based on miRNA biogenesis model. Its mapper module aligns reads to the genome, further its miRDeep2 module excises potential hairpin structures and assigns a probabilistic score to decide if hairpins are true precursors. Quantifier module of miRDeep2 is responsible for quantification of miRNA expression (Friedlander *et al.*, 2012). This analysis indicated that there are 44, 40 and 142 novel miRNAs expressed in liver, pectoralis major muscle and hypothalamus of HWS, respectively. Whereas 47, 47, and 164 novel miRNAs showed expression in liver, pectoralis major muscle and hypothalamus of LWS, respectively. Likewise, liver, pectoralis major muscle and hypothalamus tissues showed expression of 166, 183 and 256 known miRNAs in HWS and 177, 177, and 250 known miRNAs in LWS, respectively. These results suggest that hypothalamus expresses a larger number of miRNAs in both growth lines compared to liver and pectoralis major muscle. Hence, our data is revealing larger number of novel and known miRNAs from a larger population of 24 birds than previously appreciated.

Differential expression analysis between HWS and LWS birds was carried out in DEseq2 (Love *et al.*, 2014) that unveiled differential expression of 43, 47 and 59 miRNAs in hypothalamus, liver and pectoralis major muscle. These

data showed that there are slightly more miRNAs expressed at higher levels in HWS compared with LWS chickens. Evolutionary conservation analysis showed that 37%, 30% and 59% miRNAs differentially expressed in hypothalamus, liver and pectoralis major muscle, respectively, were highly conserved in vertebrates. Among these differentially expressed genes, there were eight miRNA genes that have a widespread expression in all three tissues. These eight DE miRNAs exhibited same expression pattern in the three tissues analysed. Functions of some of the DE-miRNAs; *miR-142*, *miR-147* and *miR-155* are established to regulate proliferation and differentiation, immune responses and adipogenesis, respectively. *miR-30d* a miRNA which is expressed in lower amounts in HWS, is known to repress cell proliferation (Wu *et al.*, 2013). In light of these findings, HWS increased expression of *miR-142*, *miR-147* and *miR-155* and HWS decreased expression of *miR-30d* may directly regulate the growth in birds. To date, functional role of *miR-1684a-3p, miR-1684b-3p* and *miR-1736* are not established. Strikingly, none of these eight DE- miRNAs has been previously reported to influence chicken skeletal muscle growth.

Earlier studies indicated that growth QTLs and selective sweep regions influencing the growth of HWS and LWS birds. Therefore, identified DE miRNAs were overlapped with known QTLs and selective sweep regions. There were a total of 13 DE-miRNAs that correlated with growth QTLs and 20 DE-miRNAs showed an overlap with selective sweep regions. There are total 10 evolutionary conserved miRNAs that are overlapping with QTL and selective sweep regions. Among these *miR-181, miR-15* and *miR-16* have established roles in skeletal muscle differentiation, apoptosis and tumor development.

Generally miRNA pairs with 3'UTR of mRNA to direct mRNA degradation. To gain insight into which mRNAs were targeted by the miRNAs we used TargetScan (Friedman *et al.*, 2009; Grimson *et al.*, 2007; Lewis *et al.*, 2005). TargetScan uses two algorithms pCT and cotext_score to detect targets of miRNAs and we have employed both of these tools for target identification. pCT is the probability of conserved target and it assigns score on the basis of miRNA seed matching with the site and by considering each individual site's phylogenetic conservation. Seed matches are always not sufficient for detectable repression so target recognition can be done by context features that measures efficacy on basis of various biochemical properties including site type, local AU rich nucleotide composition, positioning in 3'UTR and additional 3' pairing. miRNAs with higher expression level in HWS chickens versus LWS chicken were shown to target 2896 mRNAs while miRNAs with lower expression levels in HWS versus LWS targeted 948 mRNAs. Functional

annotations of targets of miRNAs with higher expression in all three tissues exhibited enrichment for various pathways including actin cytoskeleton, MAPK, focal adhesion and insulin signaling and also showed enrichment for many processes including biological, metabolic and cellular processes and gene expression. Whereas miRNAs with low expression level in hypothalamus stimulated mTOR signalling and insulin signalling in liver and pectoralis major muscle. These pathways are important for cell motility, cell survival, proliferation, differentiation, development, cellular metabolism, growth and survival. Altogether our analysis expands knowledge of chicken miRNA repertoire and provides opportunity to understand contribution of miRNAs in complex traits like growth. Our study adds new players that may contribute to establish a connection between genotype and phenotype of chicken growth selection lines.

3.4.3 Future perspectives

Recent studies illuminate that a large proportion of miRNA genes shows polycistronic transcription, this clustering harbours evolutionary and functional importance. Therefore, there is a need to identify clusters of miRNA to explain broader effects mediated by miRNAs on chicken growth. We have identified novel miRNAs expressed in three tissues but functional characterization and annotations of novel miRNAs are required to strengthen their position and role in chicken biology. Real-time PCR will be important to validate differential expression of commonly expressed miRNAs and DE-miRNAs overlapping with growth QTL and selective sweeps. Functional studies to confirm the effect of specific miRNAs to influence mRNA degradation of putative mRNAs will be crucial. Another interesting task would be to investigate the genetic polymorphism in miRNAs target mRNAs in HWS and LWS birds. In particular, such genetic variation between the HWS and LWS lines in the 3'UTR of target mRNAs may be involved in differential capacity to degrade such mRNAs. Another way to obtain insights into functional role of DE-miRNA is to generate DE-mRNA obtained from RNA-seq data of these growth chicken lines, association of DE-miRNA with DE-mRNA will facilitate to completely understand the contribution of miRNAs in growth selected chickens. To understand regulatory networks stimulated by miRNAs, there is also need to validate miRNAs target mRNAs by performing functional analysis.

4 Conclusions

This thesis provides important insights into diverse gene regulatory components including transcription factors, DNA methylation and miRNAs. In the first study we unveiled the evolutionary origins of an important transcription factor family named *ZBED* gene family. Our study unravelled that *hAT* DNA transposons were utilised to create *ZBED* family of diverse vertebrate functions and we also identified a new member of this family *C7ORF29* that is a closest relative of *ZBED6*. Our study reinforces important role of DNA transposons in the evolution of genomes by giving birth to the important gene regulators of diverse host functions. Second study demonstrated that hypermethylation of ZBED6 binding regions led to transcriptional silencing of corresponding genes in colon adenocarcinoma. These hypermethylated genes contain binding sites of another transcription factor PAX5 in vicinity of ZBED6 binding sites. In third study we performed functional studies to elucidate functional role of ZBED6 and showed that ZBED6 is a repressor for *IGF2* transcription and an activator for *Twist2* transcription. We also found that ZBED6 was strongly associated with histone marks of active promoter rather than repressive histone marks. Fourth study shed light on expression of both known and novel miRNAs in weight selection lines of chicken. We found differentially expressed miRNAs in hypothalamus, liver and muscle. Many of these miRNAs are evolutionary conserved and overlap with growth QTLs and selective sweep regions and also participate in important biological processes and pathways influencing the growth. In short, these studies expand knowledge about a novel transcription factor ZBED6 and also shed light on the role of miRNAs in complex trait of growth. Notably, this thesis also indicates the importance and strength of bioinformatics analyses of high throughput data to decipher the gene regulatory mechanisms. These outcomes will assist to understand the complexity of gene regulatory networks in vertebrates and will also provide basis for future investigations.

References

Andreote, A.P., Rosario, M.F., Ledur, M.C., Jorge, E.C., Sonstegard, T.S., Matukumalli, L. & Coutinho, L.L. (2014). Identification and characterization of microRNAs expressed in chicken skeletal muscle. *Genet Mol Res,* 13(1), pp. 1465-79.

Aravind, L. (2000). The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem Sci,* 25(9), pp. 421-3.

Arensburger, P., Hice, R.H., Zhou, L., Smith, R.C., Tom, A.C., Wright, J.A., Knapp, J., O'Brochta, D.A., Craig, N.L. & Atkinson, P.W. (2011). Phylogenetic and functional characterization of the hAT transposon superfamily. *Genetics,* 188(1), pp. 45-57.

Ball, M.P., Li, J.B., Gao, Y., Lee, J.H., LeProust, E.M., Park, I.H., Xie, B., Daley, G.Q. & Church, G.M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol,* 27(4), pp. 361-8.

Baylin, S.B. & Ohm, J.E. (2006). Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer,* 6(2), pp. 107-16.

Bazzini, A.A., Lee, M.T. & Giraldez, A.J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science,* 336(6078), pp. 233-7.

Boehm, M. & Slack, F. (2005). A developmental timing microRNA and its target regulate life span in C. elegans. *Science,* 310(5756), pp. 1954-7.

Braunschweig, M.H., Van Laere, A.S., Buys, N., Andersson, L. & Andersson, G. (2004). IGF2 antisense transcript expression in porcine postnatal muscle is affected by a quantitative trait nucleotide in intron 3. *Genomics,* 84(6), pp. 1021-9.

Burkhart, C.A., Cherry, J.A., Van Krey, H.P. & Siegel, P.B. (1983). Genetic selection for growth rate alters hypothalamic satiety mechanisms in chickens. *Behav Genet,* 13(3), pp. 295-300.

Calvi, B.R., Hong, T.J., Findley, S.D. & Gelbart, W.M. (1991). Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: hobo, Activator, and Tam3. *Cell,* 66(3), pp. 465-71.

Carlborg, O., Jacobsson, L., Ahgren, P., Siegel, P. & Andersson, L. (2006). Epistasis and the release of genetic variation during long-term selection. *Nat Genet,* 38(4), pp. 418-20.

Casola, C., Hucks, D. & Feschotte, C. (2008). Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol Biol Evol,* 25(1), pp. 29-41.

Chen, T., Li, M., Ding, Y., Zhang, L.S., Xi, Y., Pan, W.J., Tao, D.L., Wang, J.Y. & Li, L. (2009). Identification of zinc-finger BED domain-containing 3 (Zbed3) as a novel Axin-interacting protein that activates Wnt/beta-catenin signaling. *J Biol Chem,* 284(11), pp. 6683-9.

Colussi, D., Brandi, G., Bazzoli, F. & Ricciardiello, L. (2013). Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention. *Int J Mol Sci,* 14(8), pp. 16365-85.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigo, R. & Gingeras, T.R. (2012). Landscape of transcription in human cells. *Nature,* 489(7414), pp. 101-8.

Djuranovic, S., Nahvi, A. & Green, R. (2012). miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science,* 336(6078), pp. 237-40.

Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol,* 29(8), pp. 1969-73.

Dunnington, E.A. & Siegel, P.B. (1996). Long-term divergent selection for eight-week body weight in white Plymouth rock chickens. *Poult Sci,* 75(10), pp. 1168-79.

Eriksson, J., Larson, G., Gunnarsson, U., Bed'hom, B., Tixier-Boichard, M., Stromstedt, L., Wright, D., Jungerius, A., Vereijken, A., Randi, E., Jensen, P. & Andersson, L. (2008). Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. *PLoS Genet,* 4(2), p. e1000010.

Eulalio, A., Huntzinger, E., Nishihara, T., Rehwinkel, J., Fauser, M. & Izaurralde, E. (2009). Deadenylation is a widespread effect of miRNA regulation. *Rna,* 15(1), pp. 21-32.

Farber, D.B., Theendakara, V.P., Akhmedov, N.B. & Saghizadeh, M. (2010). ZBED4, a novel retinal protein expressed in cones and Muller cells. *Adv Exp Med Biol,* 664, pp. 79-87.

Feng, S., Jacobsen, S.E. & Reik, W. (2010). Epigenetic reprogramming in plant and animal development. *Science,* 330(6004), pp. 622-7.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet,* 9(5), pp. 397-405.

Feschotte, C. & Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet,* 41, pp. 331-68.

Filipowicz, W., Bhattacharyya, S.N. & Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet,* 9(2), pp. 102-14.

Friedlander, M.R., Mackowiak, S.D., Li, N., Chen, W. & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res,* 40(1), pp. 37-52.

Friedman, R.C., Farh, K.K., Burge, C.B. & Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res,* 19(1), pp. 92-105.

Fumihito, A., Miyake, T., Takada, M., Shingu, R., Endo, T., Gojobori, T., Kondo, N. & Ohno, S. (1996). Monophyletic origin and unique dispersal patterns of domestic fowls. *Proc Natl Acad Sci U S A,* 93(13), pp. 6792-5.

Gaston, K. & Jayaraman, P.S. (2003). Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cell Mol Life Sci,* 60(4), pp. 721-41.

Gebhard, C., Benner, C., Ehrich, M., Schwarzfischer, L., Schilling, E., Klug, M., Dietmaier, W., Thiede, C., Holler, E., Andreesen, R. & Rehli, M. (2010). General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells. *Cancer Res,* 70(4), pp. 1398-407.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A.P., Cayting, P., Charos, A., Chen, D.Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E.C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T.E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K.Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P.J., Myers, R.M., Weissman, S.M. & Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature,* 489(7414), pp. 91-100.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. & Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell,* 27(1), pp. 91-105.

Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R.A. & Feinberg, A.P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nat Genet,* 43(8), pp. 768-75.

Hickman, A.B., Perez, Z.N., Zhou, L., Musingarimi, P., Ghirlando, R., Hinshaw, J.E., Craig, N.L. & Dyda, F. (2005). Molecular architecture of a eukaryotic DNA transposase. *Nat Struct Mol Biol,* 12(8), pp. 715-21.

Hicks, J.A., Tembhurne, P. & Liu, H.C. (2008). MicroRNA expression in chicken embryos. *Poult Sci,* 87(11), pp. 2335-43.

Ho Sui, S.J., Fulton, D.L., Arenillas, D.J., Kwon, A.T. & Wasserman, W.W. (2007). oPOSSUM: integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res,* 35(Web Server issue), pp. W245-52.

Ho Sui, S.J., Mortimer, J.R., Arenillas, D.J., Brumm, J., Walsh, C.J., Kennedy, B.P. & Wasserman, W.W. (2005). oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res,* 33(10), pp. 3154-64.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H.N., Shin, J., Cox, E., Rho, H.S., Woodard, C., Xia, S., Liu, S., Lyu, H., Ming, G.L., Wade, H., Song, H., Qian, J. & Zhu, H. (2013). DNA methylation presents distinct binding sites for human transcription factors. *Elife,* 2, p. e00726.

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J.B., Sabunciyan, S. & Feinberg, A.P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet,* 41(2), pp. 178-86.

Jacobsson, L., Park, H.B., Wahlberg, P., Fredriksson, R., Perez-Enciso, M., Siegel, P.B. & Andersson, L. (2005). Many QTLs with minor additive effects are associated with a large difference in growth between two selection lines in chickens. *Genet Res,* 86(2), pp. 115-25.

Jeon, J.T., Carlborg, O., Tornsten, A., Giuffra, E., Amarger, V., Chardon, P., Andersson-Eklund, L., Andersson, K., Hansson, I., Lundstrom, K. & Andersson, L. (1999). A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nat Genet,* 21(2), pp. 157-8.

Jia, Y., Yuan, L., Hu, W., Luo, Y., Suo, L., Yang, M., Chen, S., Wang, Y., Liu, H., Yang, G. & Li, L. (2014). Zinc-finger BED domain-containing 3 (Zbed3) is a novel secreted protein associated with insulin resistance in humans. *J Intern Med,* 275(5), pp. 522-33.

Jiang, L., Wallerman, O., Younis, S., Rubin, C.J., Gilbert, E.R., Sundstrom, E., Ghazal, A., Zhang, X., Wang, L., Mikkelsen, T.S., Andersson, G. & Andersson, L. (2014). ZBED6 modulates the transcription of myogenic genes in mouse myoblast cells. *PLoS One,* 9(4), p. e94187.

Johansson, A.M., Pettersson, M.E., Siegel, P.B. & Carlborg, O. (2010). Genome-wide effects of long-term divergent selection. *PLoS Genet,* 6(11), p. e1001188.

Jordan, I.K., Rogozin, I.B., Glazko, G.V. & Koonin, E.V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet,* 19(2), pp. 68-72.

Ka, S., Albert, F.W., Denbow, D.M., Paabo, S., Siegel, P.B., Andersson, L. & Hallbook, F. (2011). Differentially expressed genes in hypothalamus in relation to genomic regions under selection in two chicken lines resulting from divergent selection for high or low body weight. *Neurogenetics,* 12(3), pp. 211-21.

Kanginakudru, S., Metta, M., Jakati, R.D. & Nagaraju, J. (2008). Genetic evidence from Indian red jungle fowl corroborates multiple domestication of modern day chicken. *BMC Evol Biol,* 8, p. 174.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe,

S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J. & International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature,* 409(6822), pp. 860-921.

Latchman, D.S. (2008). *Eukaryotic transcription factors*. 5th. ed. Amsterdam ; London: Elsevier/Academic Press.

Lee, R.C., Feinbaum, R.L. & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75(5), pp. 843-54.

Levine, M. & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature,* 424(6945), pp. 147-51.

Lewis, B.P., Burge, C.B. & Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell,* 120(1), pp. 15-20.

Li, T., Wu, R., Zhang, Y. & Zhu, D. (2011). A systematic analysis of the skeletal muscle miRNA transcriptome of chicken varieties with divergent skeletal muscle growth identifies novel miRNAs and differentially expressed miRNAs. *BMC Genomics,* 12, p. 186.

Li, X. & Carthew, R.W. (2005). A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye. *Cell,* 123(7), pp. 1267-77.

Lin, S., Li, H., Mu, H., Luo, W., Li, Y., Jia, X., Wang, S., Jia, X., Nie, Q., Li, Y. & Zhang, X. (2012). Let-7b regulates the expression of the growth hormone receptor gene in deletion-type dwarf chickens. *BMC Genomics,* 13, p. 306.

Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E. & Slack, F.J. (2003). The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. *Dev Cell,* 4(5), pp. 639-50.

Love, M.I., Huber, W. & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*.

Lowe, C.B. & Haussler, D. (2012). 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS One,* 7(8), p. e43128.

Markljung, E., Jiang, L., Jaffe, J.D., Mikkelsen, T.S., Wallerman, O., Larhammar, M., Zhang, X., Wang, L., Saenz-Vash, V., Gnirke, A., Lindroth, A.M., Barres, R., Yan, J., Stromberg, S., De, S., Ponten, F., Lander, E.S., Carr, S.A., Zierath, J.R., Kullander, K., Wadelius, C., Lindblad-Toh, K., Andersson, G., Hjalm, G. & Andersson, L. (2009). ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. *PLoS Biol,* 7(12), p. e1000256.

Marquez, G.C., Siegel, P.B. & Lewis, R.M. (2010). Genetic diversity and population structure in lines of chickens divergently selected for high and low 8-week body weight. *Poult Sci,* 89(12), pp. 2580-8.

Maston, G.A., Evans, S.K. & Green, M.R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet,* 7, pp. 29-59.

McClintock, B. (1953). Induction of Instability at Selected Loci in Maize. *Genetics,* 38(6), pp. 579-99.

McGhee, J.D. & Ginder, G.D. (1979). Specific DNA methylation sites in the vicinity of the chicken beta-globin genes. *Nature,* 280(5721), pp. 419-20.

Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., Hu, H., Khaitovich, P. & Kaessmann, H. (2013). Birth and expression evolution of mammalian microRNA genes. *Genome Res,* 23(1), pp. 34-45.

Miller, W.J., McDonald, J.F. & Pinsker, W. (1997). Molecular domestication of mobile elements. *Genetica,* 100(1-3), pp. 261-70.

Mokhonov, V.V., Theendakara, V.P., Gribanova, Y.E., Ahmedli, N.B. & Farber, D.B. (2012). Sequence-specific binding of recombinant Zbed4 to DNA: insights into Zbed4 participation in gene transcription and its association with other proteins. *PLoS One,* 7(5), p. e35317.

Morozova, N., Zinovyev, A., Nonne, N., Pritchard, L.L., Gorban, A.N. & Harel-Bellan, A. (2012). Kinetic signatures of microRNA modes of action. *Rna,* 18(9), pp. 1635-55.

Nezer, C., Moreau, L., Brouwers, B., Coppieters, W., Detilleux, J., Hanset, R., Karim, L., Kvasz, A., Leroy, P. & Georges, M. (1999). An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nat Genet,* 21(2), pp. 155-6.

Pabo, C.O. & Sauer, R.T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem,* 61, pp. 1053-95.

Pace, J.K., 2nd & Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res,* 17(4), pp. 422-32.

Pettersson, M.E., Johansson, A.M., Siegel, P.B. & Carlborg, O. (2013). Dynamics of adaptive alleles in divergently selected body weight lines of chickens. *G3 (Bethesda),* 3(12), pp. 2305-12.

Ray, D.A., Feschotte, C., Pagan, H.J., Smith, J.D., Pritham, E.J., Arensburger, P., Atkinson, P.W. & Craig, N.L. (2008). Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Res,* 18(5), pp. 717-28.

Rhee, I., Bachman, K.E., Park, B.H., Jair, K.W., Yen, R.W., Schuebel, K.E., Cui, H., Feinberg, A.P., Lengauer, C., Kinzler, K.W., Baylin, S.B. & Vogelstein, B. (2002). DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature,* 416(6880), pp. 552-6.

Ronquist, F. & Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19(12), pp. 1572-4.

Saghizadeh, M., Akhmedov, N.B., Yamashita, C.K., Gribanova, Y., Theendakara, V., Mendoza, E., Nelson, S.F., Ljubimov, A.V. & Farber, D.B. (2009). ZBED4, a BED-type zinc-finger protein in the cones of the human retina. *Invest Ophthalmol Vis Sci,* 50(8), pp. 3580-8.

Siegel, P.B., Dodgson, J.B. & Andersson, L. (2006). Progress from chicken genetics to the chicken genome. *Poult Sci,* 85(12), pp. 2050-60.

Singal, R. & Ginder, G.D. (1999). DNA methylation. *Blood,* 93(12), pp. 4059-70.

Sinzelle, L., Izsvak, Z. & Ivics, Z. (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci,* 66(6), pp. 1073-93.

Sokol, N.S. & Ambros, V. (2005). Mesodermally expressed Drosophila microRNA-1 is regulated by Twist and is required in muscles during larval growth. *Genes Dev,* 19(19), pp. 2343-54.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., Tiwari, V.K. & Schubeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature,* 480(7378), pp. 490-5.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics,* 22(21), pp. 2688-90.

Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z. & Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol,* 16(5), pp. 564-71.

Suzuki, M.M. & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet,* 9(6), pp. 465-76.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. & Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res,* 25(24), pp. 4876-82.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul,

R., Furey, T.S., Dekker, J., Crawford, G.E. & Stamatoyannopoulos, J.A. (2012). The accessible chromatin landscape of the human genome. *Nature,* 489(7414), pp. 75-82.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J.G., Baylin, S.B. & Issa, J.P. (1999). CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A,* 96(15), pp. 8681-6.

Van Laere, A.S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., Archibald, A.L., Haley, C.S., Buys, N., Tally, M., Andersson, G., Georges, M. & Andersson, L. (2003). A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature,* 425(6960), pp. 832-6.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet,* 10(4), pp. 252-63.

Volff, J.N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays,* 28(9), pp. 913-22.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., Rando, O.J., Birney, E., Myers, R.M., Noble, W.S., Snyder, M. & Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res,* 22(9), pp. 1798-812.

Wang, X., Jiang, L., Wallerman, O., Engstrom, U., Ameur, A., Gupta, R.K., Qi, Y., Andersson, L. & Welsh, N. (2013). Transcription factor ZBED6 affects gene expression, proliferation, and cell death in pancreatic beta cells. *Proc Natl Acad Sci U S A,* 110(40), pp. 15997-6002.

Warren, W.D., Atkinson, P.W. & O'Brochta, D.A. (1994). The Hermes transposable element from the house fly, Musca domestica, is a short inverted repeat-type element of the hobo, Ac, and Tam3 (hAT) element family. *Genet Res,* 64(2), pp. 87-97.

Weisenberger, D.J., Siegmund, K.D., Campan, M., Young, J., Long, T.I., Faasse, M.A., Kang, G.H., Widschwendter, M., Weener, D., Buchanan, D., Koh, H., Simms, L., Barker, M., Leggett, B., Levine, J., Kim, M., French, A.J., Thibodeau, S.N., Jass, J., Haile, R. & Laird, P.W. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet,* 38(7), pp. 787-93.

Whitfield, T.W., Wang, J., Collins, P.J., Partridge, E.C., Aldred, S.F., Trinklein, N.D., Myers, R.M. & Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol,* 13(9), p. R50.

Wu, C., Jin, B., Chen, L., Zhuo, D., Zhang, Z., Gong, K. & Mao, Z. (2013). MiR-30d induces apoptosis and is regulated by the Akt/FOXO pathway in renal cell carcinoma. *Cell Signal,* 25(5), pp. 1212-21.

Xu, P., Vernooy, S.Y., Guo, M. & Hay, B.A. (2003). The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr Biol,* 13(9), pp. 790-5.

Yamashita, D., Sano, Y., Adachi, Y., Okamoto, Y., Osada, H., Takahashi, T., Yamaguchi, T., Osumi, T. & Hirose, F. (2007). hDREF regulates cell proliferation and expression of ribosomal protein genes. *Mol Cell Biol,* 27(6), pp. 2003-13.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24(8), pp. 1586-91.

Yao, J., Wang, Y., Wang, W., Wang, N. & Li, H. (2011). Solexa sequencing analysis of chicken pre-adipocyte microRNAs. *Biosci Biotechnol Biochem,* 75(1), pp. 54-61.

Yoo, A.S. & Greenwald, I. (2005). LIN-12/Notch activation leads to microRNA-mediated down-regulation of Vav in C. elegans. *Science,* 310(5752), pp. 1330-3.

Zhao, Y., Samal, E. & Srivastava, D. (2005). Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature,* 436(7048), pp. 214-20.

Zhou, L., Mitra, R., Atkinson, P.W., Hickman, A.B., Dyda, F. & Craig, N.L. (2004). Transposition of hAT elements links transposable elements and V(D)J recombination. *Nature,* 432(7020), pp. 995-1001.

Acknowledgements

This work was performed at SLU department of Animal Breeding and Genetics, Uppsala, Sweden. During the period of my PhD studies, many people have contributed. I would like to thank all of them.

First of all, I would like to express my deepest appreciation and gratitude to my main supervisor Göran Andersson. Without his guidance and persistent help this dissertation would not be possible. His enthusiasm, encouragement, kindness and faith in me have been extremely helpful. He always gave his time and vast knowledge generously for discussions and answering my questions. He helped me to remove all obstacles in the course of my PhD.

I feel pleasure to place on record my deep sense of thankfulness and indebtedness to my co-supervisor Leif Andersson for continually and convincingly conveying a spirit of research. His ideas helped me to improve my research capabilities.

I am honored to express my deep sense of gratitude to my other co-supervisors Erik Bongcam-Rudloff, Patric Jern, Alvaro Martinez Barrio and Anna-Maja Molin. All provided me a very good support specially Maja always took a great interest in my thesis and provided her valuable assistance in the pursuit of my research. Patric and Alex Hayward helped me in learning phylogenetics and guided me in scientific writing. It was always interesting to discuss bioinformatics with Erik and Alvaro.

My deepest gratitude and appreciations to **Higher Education Commission**, **Pakistan** for providing funds to perform my PhD research.

I want to thank all former and present members of ZBED6 group for sharing their precious data and discussing different scientific concepts during our weekly meetings.
My deepest gratitude to all Pakistani colleagues and families who made my long stay social, enjoyable, convenient and entertaining in Sweden

I would like to acknowledge very kind and polite attitude of my department administrative staff and other colleagues as well.

Finally, I want to thank my family especially three darling ladies; my grandmother, my mother and my sister for providing their invaluable love, care, support and encouragement at every step of life.