Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

Bioinformatics analysis of ZBED6, a novel transcription factor in mammals

Shumaila Sayyab

Department of Animal Breeding and Genetics
Examensarbete 331
Uppsala 2010

Master's Thesis, 30 HEC

One-Year Master's Programme in Biology
– Bioinformatics

SLU

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Bioinformatics analysis of ZBED6, a novel transcription factor in mammals

Shumaila Sayyab

Supervisors:
Göran Andersson, SLU, Department of Animal Breeding and Genetics
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics
Examiner:
Gabriella Lindgren, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: BI1021
Programme: One-Year Master's Programme in Biology
- Bioinformatics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2010
Name of series: Examensarbete 331
Department of Animal Breeding and Genetics, SLU
On-line publication: http://epsilon.slu.se

Key words: Chromatin Immuno precipitation ChIP sequencing; non-coding RNA; transcription factor; overrepresentation; enhancer elements; co regulation; repressor.

Table of Contents

ABSTRACT

The identification of a regulatory mutation in the intron 3 of Insulin like growth factor 2, was a major finding. This single nucleotide mutation in the non coding region abrogates the interaction of nuclear factor ZBED6, resulting in the 3-4% increase in the muscle mass of domesticated pigs. The mutation was observed in the evolutionary conserved CpG island that is hypomethylated in skeletal muscle. ZBED6 has been derived from a domesticated DNA transposon exclusively present in the placental mammals. Chromatin immuno precipitation (ChIP) sequencing in mouse C2C12 myoblasts using anti-ZBED6 antibody identified 2499 ZBED6 binding fragments. The de novo search on the binding fragments of ZBED6 showed the consensus sequence of 5′-GCTCG-3′. ZBED6 binding fragments contain more than 1200 genes with annotated functions in biological processes, transcriptional regulation, neurogenesis, cell signaling and muscle development. In present study we have done bioinformatics analysis on the ZBED6 using TRANSFAC database professional version 2010.1 in order to identify the other transcription factors co-regulating the expression of ZBED6 target genes. The ChIP data (ZBED6 target genes) and microarray expression data (siRNA silenced ZBED6) in mouse C2C12 cells were used in this study for finding the binding sites for transcription factors in the promoter regions. The genes associated with ZBED6 showed significant overrepresentation of binding sites of transcription factors SP1, ZF5, E2F1, ZBED6, AP2alpha and KROX in their promoter regions. Majority of factors found have GC rich binding sites and belongs to zinc finger families. The obtained factors show role in tumor suppression. The microarray expression data analysis showed that MEF2 and SRF transcription factors binding sites are significantly present in the promoters of co-expressed genes. The ZBED6 binding sites that were at a distance of 500 kb away from known transcription start site TSS, showed OCT1 and IRF1 binding sites. There is a possibility that these factors are the enhancer elements for many ZBED6 target genes. Few long non-coding RNAs were also identified in the vicinity of ZBED6 binding sites present at a distance of 500 kb away from known TSS.

INTRODUCTION

Muscle development in domesticated mammals

The characteristics of meat (with low fat contents and high muscle fiber) are equally important in the production of meat and the consumer favor the lean and tender meat. The muscle development is identified in many species of domesticated mammals (e.g. cattle, sheep, pig and chicken). The pig is used for meat production throughout the world except for most Muslim countries. The history of pig domestication dates back at ~9000 years ago in the Near East and analyses of mitochondrial sequence showed that it occurred at multiple locations throughout Europe and Asia (Giuffra et al., 2000). The ancestor of domesticated pigs is European wild boar (*Sus scrofa*) (Larson et al., 2005). The selection on lean meat during the last 60 years in domestic pigs resulted in increased muscle mass and low fat content (Markljung et al., 2009). Present day, breeds like Large White, Duroc, Landrace, Pietrain and Hampshire are used to produce the lean meat with the characteristics in terms of tenderness, leanness and moisture (Chen et al., 2007). As a result of this selection, the pig breeds with favorable muscle growth allele variants have increased with a concomitant reduced frequency of fat content allele variants. Genetic studies aimed at identifying loci influencing muscle growth have identified new genes and factors that might be involved in the regulation of muscle development and differentiation process.

Three different genes have been identified to have mutations affecting muscle growth; these include missense mutaion in the *PRKAG3* gene that plays a key role in the regulation of energy metabolism in skeletal muscle, mutation in ryanodine receptor gene *RYR1* that causes the malignant hyperthermia, which is a recessive disorder. This gene encodes the ryanodine receptor found in skeletal muscle. Furthermore, a single nucleotide substitution in the insulin-like growth factor 2 (*IGF2)* gene that is affecting the postnatal muscle growth (Milan et al., 2000; Fujii et al., 1991; Van Laere et al., 2003).

IGFs in muscle development

The muscle growth and development of meat in animals is a complex and highly regulated process. The insulin-like growth factors IGFs *i.e*. IGF1 and IGF2 are strong regulators of cellular and tissue growth. As the control of muscle growth is at the molecular, cellular and tissue level, it is critically important to understand role of IGFs in muscle development. IGFs are structurally similar polypeptides with diverse biological functions and properties. IGFs function in both endocrine and paracrine processes (Oksbjerg et al., 2004; Kokta et al., 2004). Both IGF1 and 1GF2 show great role in the proliferation and differentiation of muscle cells (Oksbjerg et al., 2004). IGF1 has been identified to be important in both proliferation and differentiation of myoblasts (Engert et al., 1996). Similarly, it has been observed that IGF2 is important for the transition from proliferation to differentiation of myoblasts (Florini et al., 1991).

Different *in vivo* knockdown mice studies showed great importance of IGF1 and IGF2 as it reduces the number of cells in different tissues including muscle cells (Liu et al., 1993). The expression level of *IGF2* mRNA increases and reaches maximum expression level at the stage of myogenesis in growing muscle fibers whereas the level of *IGF1* mRNA expression increases slightly at initial stages but in neonatal pigs it increases at maximum point (Gerrard et al., 1998). The growth hormone (GH) stimulates these factors (Oksbjerg et al., 2004). The IGFs interacts with three different receptors i.e. IGF1R, IGF2R and insulin receptor. The IGF1 can bind with IGF1R and insulin receptor while IGF2 can bind with IGF1R and IGF2R with different binding affinities. IGF1 is the natural activator of AKT signaling pathway and inhibitor of programmed cell death. The signaling pathways that mediate the effects of IGFs in the muscle cells are Phosphatidylinositol-3-kinase (PI3K) pathway and mitogen activated protein kinase (MAPK) pathway. It has been defined that IGF1-induced MAPK pathway is mainly involved in proliferation while PI3K pathway is involved in cell differentiation (Coolican et al., 1997). On the other hand, IGF2 has been shown to induce differentiation through PI3K pathway (Kaliman et al., 1999). The biological activities of these factors are greatly regulated by the family of IGF binding proteins (IGFBPs). These vertebrate IGFBPs belong to a group of secreted proteins that binds to IGF1 and IGF2 and control their actions. Both circulating and locally expressed IGF1 has been shown to be involved in postnatal muscle growth in many studies (Oksbjerg et al., 2004). Not many studies indicate the effect of IGF2 on the postnatal muscle growth although some indicate its effect on the fat deposition (Owens et al., 1999). Due to a point mutation in the regulatory region of the *IGF2* gene, an increase in skeletal muscle-specific *IGF2* mRNA expression in pig muscle that increased the muscle mass by 3-4% in postnatal pigs (Van Laere et al., 2003).

Gene Regulation

All the cells have same set of genes but only few genes are expressed in each cell type, this is controlled by the regulatory sequences which are present in the non protein coding regions where the regulatory proteins binds. The structure of chromosomes *i.e.* the chromatin determines the fate of the transcription mainly through the DNA methylation and various histone modifications. *Cis*-acting gene regulatory regions control the expression of genes in space, time and quantity. These include

- Promoter regions (core promoter, upstream promoter and regulatory promoter)
- Enhancer regions (binds repressors and activators)

Changes in chromatin structure (epigenetic modifications) at regulatory regions involve histone acetylation and histone methylation and on DNA, (CpG) methylation that can lead to the activation or deactivation of certain transcriptional processes at different time and place. For gene expression the regulatory sequence must be accessible to the RNA polymerases and other transcription factors.

Regulatory Mutation in *IGF2*

Regulatory mutations influence transcription and occur in the regulatory regions (promoters or enhancers) of the genes. The mutations may influence binding of the regulatory transcription factors and preventing their binding to that regulatory region. Quantitative Trait Locus (QTL) mapping is the study of Quantitative Traits (a complex trait that can be measured or continuous trait *e.g.* body weight) that is affected by the genotype and environment. The objective of QTL mapping study is to identify a region or regions in the genome that are associated with some particular complex trait including a complex disease. Ultimately, detailed fine-mapping and re-sequencing may allow the identification of causative mutation(s) for the QTL.

A paternally expressed imprinted QTL that is affecting the skeletal and cardiac muscle growth and fat deposition was first identified in a study using an intercross between the European wild boar with large white domestic pigs and also Pietrain with large white pigs and the imprinted QTL mapped to the *IGF2* locus in pigs (Jeon et al., 1999; Nezer et al., 1999). This intercross allowed QTL mapping of genes influencing increased muscle mass and reduced fat deposition. In a subsequent follow-up study, the haplotype sharing approach was used and it was found that the favorable alleles had gone through a selective sweep as a result of strong selection for lean meat. A single nucleotide point mutation denoted Q was identified in the QTL bearing region, this mutation was not present in the wild type allele q (within the 250 kb interval between the markers 370SNP6/15 and SWC9) where this region contain the *IGF2* and *INS* genes. The *IGF2* gene was identified as an excellent candidate gene because it is paternally expressed (Van Laere et al., 2003). The *IGF2* gene was resequenced from multiple q and Q haplotypes and the mutated Q haplotype was shown to harbor a transition mutation Guanine to Adenine (G to A) in intron 3 at nucleotide 3072, which is the causative Quantitative trait nucleotide QTN (Van Laere et al., 2003). This QTN is located in an evolutionary conserved CpG island located between differentially methylated region 1 DMR1 and matrix attachment region (Amarger et al., 2002; Greally et al., 1997; Eden et al., 2001). IGF2 was also well-known to have critical and important function during myogenesis (Florini et al., 1995).

A region of 94 bp around the point mutation in *IGF2* intron 3 (Fig. 1) at the nucleotide 3072 is highly conserved (85 % sequence identity having 8bp palindrome upstream of the QTN) among eight species of mammals *i.e.* rat, mouse, human, rabbit, dog, horse, bovine and pig (with wild type q and mutant type Q). It also shows the other genes *INS* and *TH* along with *IGF2* within the 28.6 kb with sequence similarity within the pigs. The CpG island was found to be non-methylated at a region centered at the QTN in case of skeletal muscles while in case of liver it was methylated (Van Laere at al., 2003).

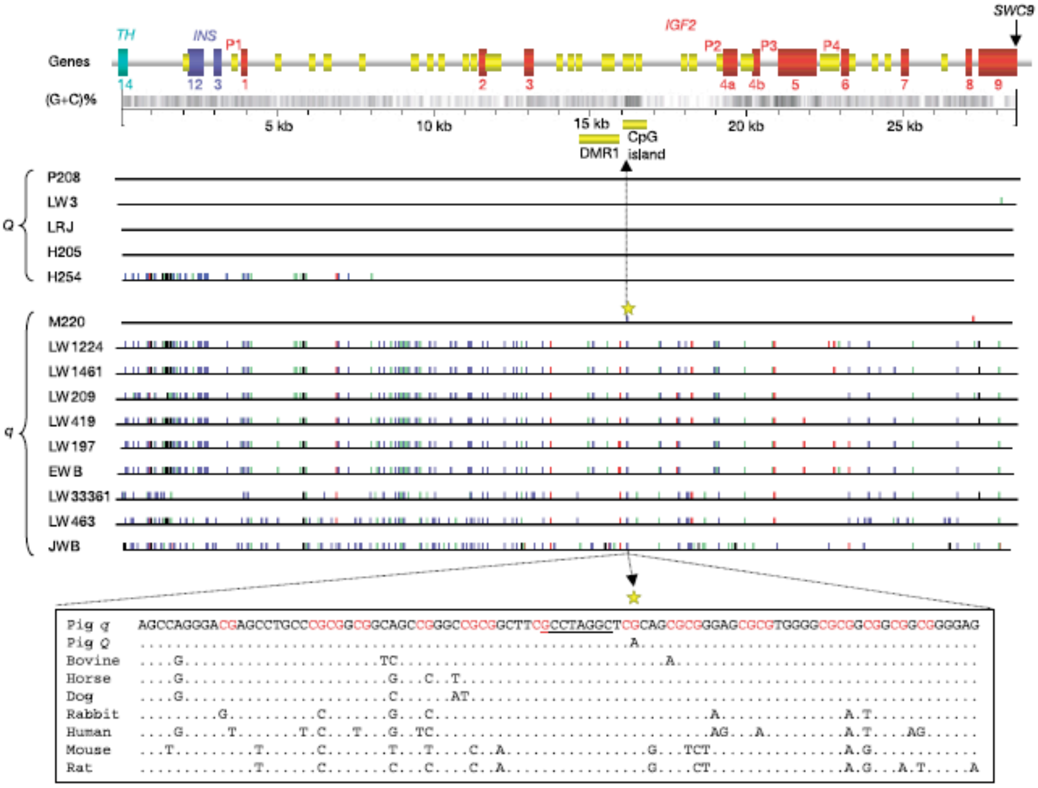

SWC9
TH
INS
IGF2
P1
P2
P3
P4
Genes
14
12
3
1
2
3
4a
4b
5
6
7
8
9
(G+C)%
5 kb
10 kb
15 kb
20 kb
25 kb
CpG island
DMR1
Q
P208
LW3
LRJ
H205
H254
q
M220
LW1224
LW1461
LW209
LW419
LW197
EWB
LW33361
LW463
JWB
Pig q AGCCAGGGACGAGCCTGCCCGCGGCGGCAGCCGGGCCGCGGCTTCGCCTAGGCTCGCAGCGCGGGAGCGCGTGGGGCGCGGCGGCGGCGGGAG
Pig Q ....A......................................
Bovine ....G.....................TC.........................A.......................................
Horse ....G......................G...C..T..
Dog ....G......................C......AT..
Rabbit .........G..........C.......G...C.............................A.............A.T...........
Human ....G.....T.......T.C...T...G..TC.............................AG...A.........A.T....AG......
Mouse ...T.........T......C.......T...T....C..A.................G...TCT..............A.G...........
Rat ............T......C.......C...C....C..A.................G....CT..............A.G...A.T.....A

Figure 1. The QTN in IGF2 intron 3 is highly conserved (Yellow star shows the position of point mutation) Van Laere et al., 2003

The biochemical and functional analyses of this mutation was performed using Electrophoretic mobility shift assay (EMSA) with double-stranded oligonucleotides corresponding to either wild type q or mutant type Q sequence. These oligonucleotides were incubated with nuclear extracts prepared from murine C2C12 myoblast cells. The study showed that the wild type q binds to a nuclear factor when it is unmethylated but cannot bind when it has mutation or methylation at the CpG dinucleotide at the QTN. Functional studies were performed using transient transfection assays using reporters driving the expression of Luciferase to determine the effect of mutation on *IGF2* transcription in myoblast C2C12 cells. Reporter constructs harboring either mutant Q or wild type q regions inserted upstream of the *IGF2* P3 promoter were constructed and used to transfect C2C12 cells. A significant increase in transcriptional activity was observed using reporter constructs with the mutant Q versus the wild type q allele. These results confirms that the wild type q sequence interacts with a repressor of *IGF2* transcription in C2C12 myoblast (Van Laere et al., 2003)

ZBED6 controlling *IGF2* expression

In order to identify the repressor of *IGF2*, the murine C2C12 myoblast cells were used and the Oligonucleotide capture and Mass Spectrometry experiments were done (Markljung et al., 2009). Two different types of nuclear extracts were made from the C2C12 cells using the stable isotope labeling of amino acids (lysine and arginine) in culture SILAC (Ong et al., 2008). The results showed the highest-fold enrichment of repressor protein (wild type q) with the previously undefined alternative splice variant of *zc3h11a* gene. The repressor was named as ZBED6.

The regulatory mutation (G to A transition) in intron 3 of *IGF2* leads to the disruption of the interaction of a repressor called ZBED6 and results in three-fold upregulation of *IGF2* mRNA expression in skeletal muscle with increase in muscle mass and meat production by 3-4% (Markljung et al., 2009). The upregulation of *IGF2* occurs in postnatal muscle but not in the fetal muscle. The pigs carrying this mutant allele at the paternal chromosome of *IGF2* showed higher expression in skeletal and cardiac muscles but not in the liver. Antisense non-coding transcript of *IGF2* expression was also found to be regulated by the same mutation (Braunschweig et al., 2004). This shows that repressor ZBED6 binding to its target sites in an unmethylated condition represses transcription of both *IGF2* and *IGF2-AS* transcription.

ZBED6 is the sixth mammalian protein with one or more BED domain. ZBED6 is derived from exapted DNA transposon that is encoded by an intron less gene located in the intron 1 of the *zch311a* gene. It contains an open reading frame of more than 900 codons, which is highly conserved among mammals (Markljung et al., 2009). It encodes a 110 kDa protein with distinct domain architecture. It contains two BED domains and a hATC dimerization domain. The BED domain was first identified in Drosophila protein BEAF and DREF that were used as the seed for the homology search using the bioinformatics analysis (Aravind; 2000). The hATC is from the hobo activator tam3. The two BED domains in ZBED6 are more closely related to each other than to any other mammalian BED domain.

ZBED6 is highly conserved among the placental mammals and its DNA binding BED domain is almost 100% conserved among 26 mammalian species (Markljung et al., 2009). The predicted *ZBED6* promoter showed that it has binding sites for transcription factors Max, Myc, Fos and NF-E2, which are playing roles in cell proliferation and cancer development. The western blot analysis of ZBED6 showed two isoforms ZBED6a and ZBED6b. It was found that ZBED6 is residing in the nucleolus of myoblast C2C12 cells that also shows its importance in the process of cell growth and proliferation (Markljung et al., 2009; Mayer et al., 2005).

ZBED6 was silenced in the C2C12 cells by using siRNA. The ZBED6-silenced and control cells were used to find the effect of silencing on the expression of specific gene. *IGF2* mRNA expression was low on the first day in both control and silenced cells. On the other hand its expression on day 6 was increased in ZBED6-silenced cells as compared to control cells. The silencing of ZBED6 showed high levels *IGF2* expression on day 6, increased cell proliferation on day 3 and wound healing on day3 which also confirms its role as repressor for *IGF2* and its functional significance (Markljung et al., 2009).

The ChIP-sequencing experiment was carried out using the AB SOLiD technology. In this experiment the C2C12 mouse cells were used with anti-ZBED6 antibody to find the target genes of ZBED6 other than *IGF2.* The results provided 24 million reads aligned to the mouse genome. 2499 peaks were found with the minimum of 15 overlapping extended reads (Markljung et al., 2009). ZBED6 peaks were often found in the near vicinity of transcription start sites TSS (with ~50% located within 5kb) and about 28% ZBED6 binding sites were in CpG Island. It was seen the ZBED6 targets about 262 genes that encodes for transcription factors. Using *de novo* search on the ZBED6 binding fragments (full set and subset based on enrichment) shows that it has the consensus sequence of 5′ -GCTCGC- 3′which is a perfect match to the QTN region in *IGF2.* The annotated genes (~1200 genes) shows association with regulation of many biological processes, transcriptional regulation, neurogenesis, morphogenesis, cell signaling and muscle development. Some of the target genes of ZBED6 showed significant association with different diseases in human. These may include development disorders, cancer, cardiovascular disease, connective tissue disorders etc. (Markljung et al., 2009).

Long non-coding RNAs

The main idea that the transcription does have roles other than producing proteins mainly comes from the analysis of locus control regions LCR and enhancers. In some cases it was seen that the transcription of LCR was needed to open the chromatin structure *e.g.* in case of Beta globin LCR, RNA polymerase II transcription opens the chromatin domains (Ashe et al., 1997; Gribnau et al., 2000).

The non-coding RNA genes are the DNA sequences that are transcribed but are not translated into protein. These include long non-coding RNAs and small non-coding RNAs which are abundantly transcribed in the mammalian genomes and are functionally important. The small non-coding RNAs include micro RNAs (miRNAs), small nucleolar RNAs (snoRNAs), small interfering RNAs (siRNAs), transfer RNA (tRNA), ribosomal RNAs (rRNAs) and piwi interacting RNAs (piRNAs). Many non-coding RNAs that interact with the general transcription factors are transcribed by RNA polymerase III. Other RNAs transcribed by RNA Polymerase III include tRNAs, snRNAs and 5S RNAs (Dieci et al., 2007).

Long non-coding RNAs (lncRNAs) represent one class of extragenic non protein coding transcripts which results from the RNA polymerase II transcription of RNA genes (De Santa et al., 2010). Long non-coding RNAs have the size more than 200 nucleotides (Mercer et al., 2009). Few known examples include *Xist*, *HOTAIR*, *Kcnq1ot1* etc. for which specific functions have been ascribed based on loss or gain of function experiments (Peny et al., 1996; Sleutels 2002; Mancini Dinardo et al., 2006). *HOTAIR* is the lncRNA that originate from the *HOXC* locus and have role in the epigenetic regulation. It represses the transcription across 40kb of the *HOXD* locus by changing the trimethylation state of chromatin (Rinn et al., 2007). Almost all the genes in the *Kcnq1* locus are maternally expressed except for the *kcnqot1* antisense lncRNA that is expressed paternally (Mitsuya et al., 1999). *Xist* which is another lncRNA is involved in the inactivation of X chromosomes in female placental mammals (Wutz et al., 2007).

The lncRNAs generally lack the strong conservation which is often related to their evidence for non functionality (Struhl et al, 2007). But on the other hand the *Xist* and *Air* although poorly conserved RNAs show functional significance and suggest that there might be some different selection pressure (Nesterova et al., 2001; Pang et al., 2006). The lncRNAs show specific spatio temporal expression in mouse brain, T-cell differentiation and embryonic stem cell differentiation

(Mercer et al, 2008; Dinger et al., 2008; Pang et al., 2009). These lncRNAs have shown to have role in transcription regulation, post transcriptional processing and regulation of chromosomal architecture (Amaral et al., 2008; Mehler et al., 2006). They also have the ability to recruit the activators and repressors of chromatin that changes the epigenetic nature of chromatin for the protein coding genes in the vicinity (Mattick et al., 2009; Amaral et al., 2008).

Different studies have shown the role of ncRNAs in diseases. In an expression analysis of tumor and normal cells, the change in expression of ncRNAs in different forms of cancer was observed. For example, ncRNA *OCC1* (overexpressed in colon cancer) was overexpressed in colon carcinoma cells (Pibouin et al., 2002). *PCGEMI* another ncRNA was correlated with increased proliferation and cell growth regulation in the prostate tumor cells (Fu et al., 2006). In lung cancer the *NEAT2* ncRNAs expression change was identified (Fu et., 2006). Similarly the mouse homologue of *NEAT2* showed high expression in hepatocellular carcinoma (Lin et al., 2007).

Transcription Factors and Biological Databases

Gene regulation is largely regulated by proteins called transcription factors (TFs) that recognize *cis*-regulatory sequences present in the promoter and enhancer regions of genes. The process of gene expression in eukaryotic cells is a complex process in which transcription factors interact with regulatory cis-acting DNA sequences called transcription factor binding sites (TFBS). This DNA-protein interaction results in activation or repression of transcription. Specific *cis*-elements are often modular with multiple TFBS on a given target gene. One of the main goals of current biological research is to build the complete regulatory network of organisms (Covert et al., 2004). In a specific pathological process, the complete understanding of the transcriptional regulation might help in the discovery of new drug targets. Many transcription factors and their binding sites on target genes have been identified using *in vitro* techniques such as DNase I footprinting and electrophoretic mobility shift assay EMSA (Rooney et al., 1995). These processes are very time consuming and costly. In order to cope with this problem several computational algorithms and bioinformatics studies are used for the discovery of known or novel regulatory elements. In this case, the study of the regulatory regions of set of co-regulated genes is done. The algorithms are made to find the overrepresented motifs in the regulatory regions of co-regulated genes (Hertz et al., 1999; Hughes et al., 2000).

Most transcription factors bind to short degenerative sequence motifs (6-12 bp) that frequently occur in the genome. This problem is solved by using computational approaches by modeling the transcription factor binding sites using the position weight matrix (PWM) to find the binding sites of transcription factors in the regulatory regions of gene sets. TRANSFAC is one of the databases commercially available that uses this approach (Matys et al., 2003). JASPAR (Sandelin et al., 2004) is also a database for the transcription factors matrix and it is publically available. These computational methods are based on over representation of binding sites in the promoters of co expressed genes. This allows the development of regulatory network of genes by understanding the potential transcription factors involved in the regulation of gene expression.

The aim of our study is to find the potential transcription factors that might be co-regulating the expression of number of genes with ZBED6, a novel transcription factor. The analysis will be done using the TRANSFAC database (professional version 2010.1).

MATERIALS AND METHODS

Chromatin Immuno precipitation (ChIP) data (ZBED6 target genes)

Chromatin Immuno Precipitation (ChIP) sequencing was done on mouse C2C12 myoblasts using the anti-ZBED6 antibody to identify targets of ZBED6 other than *IGF2* (Markljung et al., 2009). As a result of which 2499 binding fragments (ZBED6 peaks) were found with a minimum of 15 overlapping extended reads. The experiment was carried out using the AB SOLiD technology (Markljung et al., 2009). ZBED6 binding fragments from the ChIP data were used as the input set for bioinformatics analysis. We have used this input in TRANSFAC database version 2010.1, which is used for the prediction of transcription factor binding sites (TFBSs). Murine housekeeping genes (provided by TRANSFAC database version 2010.1 in ExPlain 3.0) were used as background or control set. We used the different sets of genes from the ChIP data as shown in Table 1A. The promoter window used was -1000 bp upstream and +1000 bp downstream to transcription start site (TSS).

Microarray data (expression of ZBED6-silenced genes)

Microarray expression analysis was done in mouse C2C12 cells using siRNA-silenced ZBED6 and normal cells at day 2, day 4 and both day time points. In this case the genes that showed differential expression at day 2, day 4 and both time points (p value < 0.05) were used as the input set while the genes with no significant change in expression (p value > 0.05) were used as background set. The promoter window used was between -500 bp to 100bp downstream to TSS. The datasets used in MatchTM from the microarray data are shown in Table 1B.

TRANSFAC

TRANSFAC is a commercial database that is used to detect the transcription factors that might be responsible for the transcriptional regulation of the given gene sets. The data collection for TRANSFAC started more than 20 years ago to find the interaction of factors and DNA binding sites. Many improvements and modifications have been done over the years (Knüppel et al., 1994, Wingender et al., 1996). It is now part of the BIOBASE Knowledge Library (BKL) database (http://www.biobase-international.com/index.php?id=transfac). The analysis of TRANSFAC results is now done using ExPlain version 3.0, which is a component of BKL. Selected species covered by TRANSFAC include *Homo sapiens*, *Canis lupus, Mus musculus* (several strains), *Rattus norvegicus*, *Gallus gallus*, *Saccharomyces cerevisiae* (several strains).In order to find transcription factors that might be important in the co-regulation of ZBED6 targeted genes we used the TRANSFAC database professional version 2010.1 that is commercially available for finding the transcription factors in the promoter region of ZBED6 target genes (Markljung et al., 2009). The professional version of TRANSFAC database provides access to large number of transcription factors, DNA and RNA binding sites, genes and matrices. The data for miRNAs ChIP-chip or ChIP-seq Fragments, promoter sequences of different species and references are only available in professional version of TRANSFAC.

MatchTM

The MatchTM is a web-based tool, designed for searching the potential transcription binding sites in the DNA sequences using the weight matrix search (Wingender et al., 2001). Match is integrated in the TRANSFAC professional database and uses its positional weight matrices (PWM) for the analysis of given gene set. It provides the user with variety of search modes by optimizing the matrix cut off values. In TRANSFAC, each matrix available has pre calculated three different cut-offs (Pickert et al., 1998).

- MinFP: minimum False Positive rate (over prediction error rate)
- MinFN: min False Negative rate (under prediction error rate)
- MinSUM: min SUM of both errors

A number of tissue-specific profiles are also provided by the TRANSFAC database. In addition to several pre-defined profiles provided by TRANSFAC that are used in MatchTM, the user could create his own profiles with the Match Profiler tool. It also allows the user to have its own specific profile that can be selected as a set of matrices with default or user-defined cut off values.

MatchTM takes DNA sequences or set of genes as an input and search for the potential TFBS using the PWM library (provided by TRANSFAC database). It searches for these binding sites within the promoter region of input genes. It gives the output in the form of list of potential TFBS and also the visual representation of their location in the sequences.

The algorithm that is used at the back end for searching for the TFBS is based on two scoring values:

- MSS: Matrix similarity score
- CSS: Core similarity score

These functions measure the quality of match (0.0 to 1.0 score) between the input sequence and the matrix, where 1.0 donates an exact match. The first five consecutive positions of matrix are conserved that are called the core of each matrix. The MSS is calculated using all positions of the matrix while the CSS is calculated using only the core positions. The cut off is the threshold that is used to match the matrix with the sequences. There are two cut offs used in MatchTM *i.e.* cut off for the matrix and cut off for core positions. The cut offs can be user-defined or default. If the CSS value is higher than the cut-off then it is accepted and added into the hash table and then it is extended at both the ends until it fits the matrix length. Then the MSS is checked and if it is higher than the cut-off it is given in the output table as the binding site for that factor that might be present in the sequence given (Kel et al., 2003). It is implemented in C language, which is wrapped by perl script to make it more user friendly (Kel et al., 1995). MatchTM is publicly available (http://www.gene-regulation.com/pub/programs.html#match). An advanced version of the tool called MatchTM Professional is available at (http://www.biobase.de). The professional version of MatchTM has access to professional TRANSFAC matrix library, which allows it to use large number of factors and matrices.

By adding ZBED6 matrix into the vertebrate profile (having all transcription factor matrices for vertebrates) that was already present in TRANSFAC, a new profile was made called ZBED6_h0.01 profile. The cut off value used was MinSUM that minimize the error rate of both false positive and false negative prediction.

Dataset used For MatchTM

Two types of datasets were used in MatchTM Analysis.

1. ChIP data (ZBED6 target genes)
2. Microarray data (expression of ZBED6-silenced genes)

Table 1. Datasets from ChIP data and Microarray data subdivided into gene sets. A. ChIP data have the Hitlist (complete set of ZBED6-target genes), Upstream and downstream ZBED6 targets have all the genes having ZBED6 binding upstream or downstream to TSS, respectively. Faraway target regions are the ones where ZBED6 binds with the distance from TSS greater than equal to 500kb. Highly enriched targets are those that show maximum overlapping reads in the ZBED6 peaks in ChIP results. B. Microarray data have the differentially expressed (DE) genes on day 2, day 4 and both days (ZBED6-silenced siRNA) with the p value < 0.05.

SR.	Data set used in MatchTM	Genes used	Total genes
1A.	**ChIP data**		
1)	Hitlist (All ZBED6 target genes)	1706	2499
2)	Upstream ZBED6 Targets	657	797
3)	Downstream ZBED6 Targets	1184	1405
3)	Faraway target regions	297	297
5)	Highly Enriched	889	1000
1B.	**Microarray data**		
1)	DE day 2 ZBED6-silenced (P value <0.05)	1600	1600
2)	DE day 4 ZBED6-silenced (P value <0.05)	1400	1400
3)	DE both day ZBED6-silenced (P value <0.05)	380	380

Long Non-coding RNA

Mammalian genomes show large amount of transcription that is occurring outside the mapped protein coding genes and generated large number of transcripts (Birney et al., 2007; Prasanth et al., 2007). Long non-coding RNAs lncRNAs represents one of the class of extragenic transcription products that results from RNA polymerase II transcription of RNA genes (De Santa et al., 2010).

From the ChIP experiment that was carried out in the mouse C2C12 cells (Markljung et al., 2009), out of 2499 ZBED6 binding fragments we selected 300 regions that show the peaks at a distance of 500 kb away from known TSS. Long non-coding RNAs were obtained from two datasets i.e. Ponjavic dataset generated by FANTOM consortium and Guttman dataset of lncRNAs (Ponjavic et al., 2007; Guttman et al., 2009). These non-coding RNAs were further filtered out to eliminate all protein coding genes (Carninci et al., 2005). The Ponjavic dataset obtained was filtered against the Ensemble mouse protein coding genes annotations resulting in the dataset of 2168 lncRNAs and Guttman 1408 lncRNAs (De Santa et al., 2010). In order to find the overlapping lncRNAs with the ZBED6 binding sites we used the intersect option of "Operate on genomic intervals" in the web based tool galaxy browser available online (http://main.g2.bx.psu.edu/root).

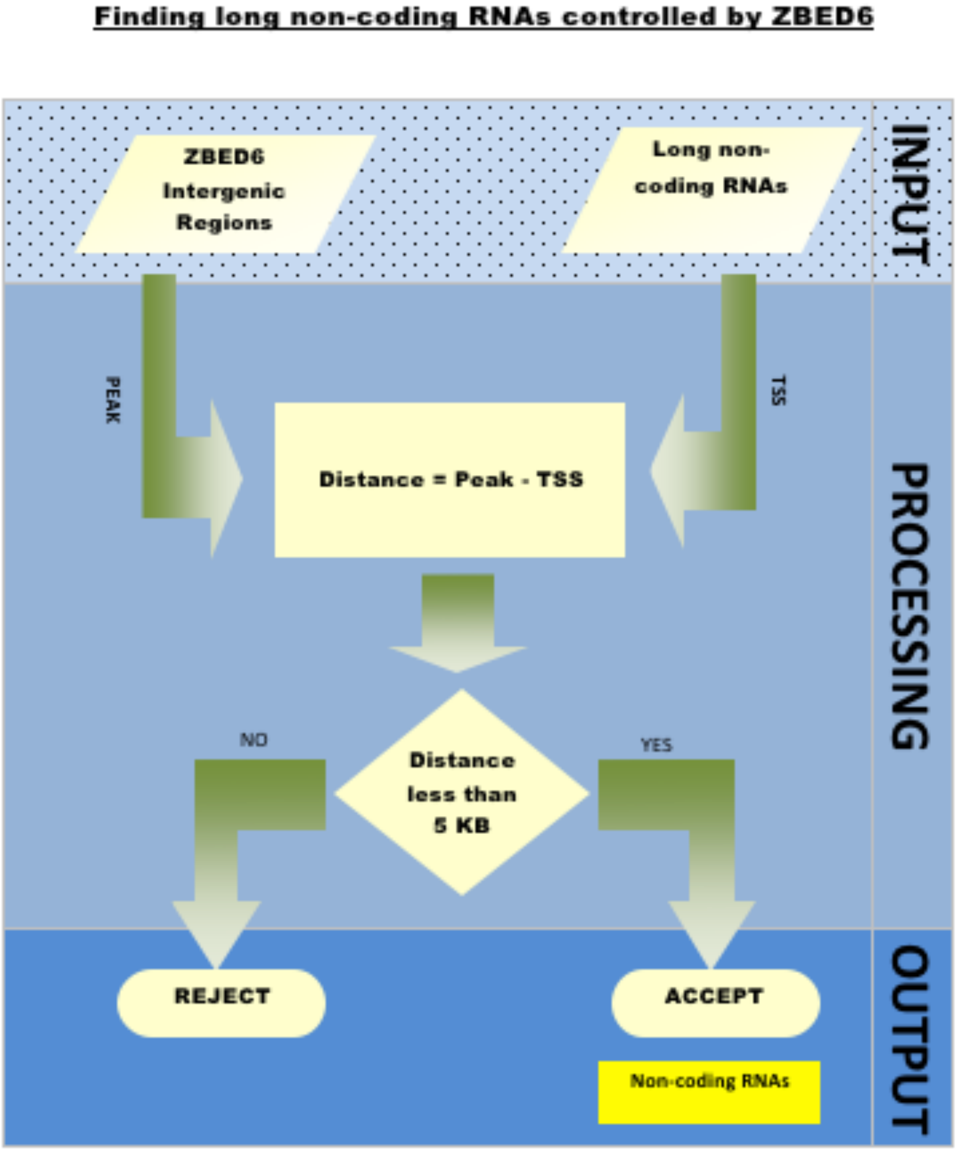


Figure 2. Flowchart to find ZBED6 peak within 5 kb of TSS of long non-coding RNAs.

We also used a perl script to find ZBED6 peaks within 5 kb of TSS of long non-coding RNA as in Figure 2. The script takes two files as an input. Here we give ZBED6 file with 300 regions (having 500 kb distance from TSS of known any known gene) and the second file is of non-coding RNAs from Ponjavic and Guttman datasets. The Distance is calculated using the formula:

Distance = Peak - TSS → Eq.1

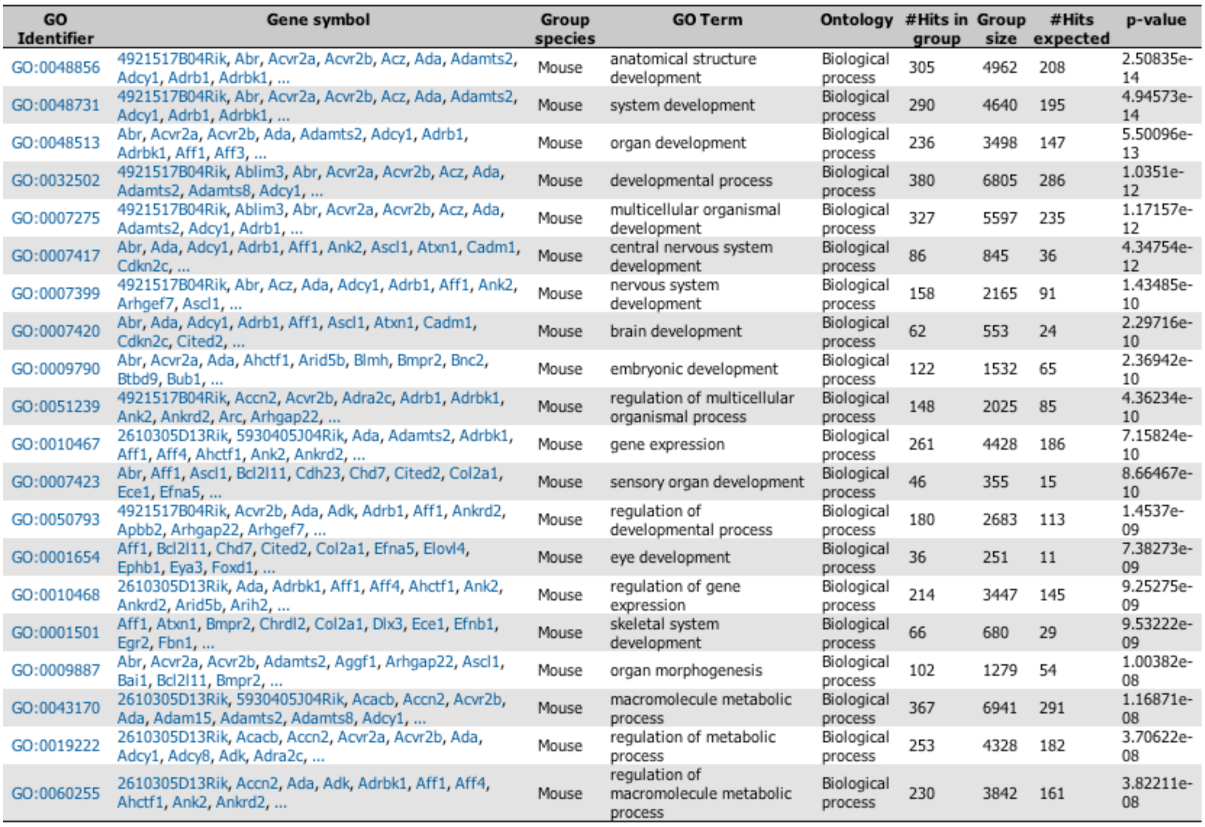
Peak in Eq.1 means binding site for ZBED6 and TSS is the transcription start site of long non-coding RNA. If the distance is less than 5 kb then the non-coding RNA and ZBED6 is shown as an output. On the other hand if distance is greater than 5 kb it is rejected. In the end it produces an output file containing all accepted ZBED6 binding sits with predicted non coding RNA start and end site, the distance of ZBED6 from lncRNAs TSS and enrichment of ZBED6. The enrichment is the overlapping read that has been obtained for each ZBED6-binding site from ChIP sequencing (Markljung et al., 2009).

RESULTS

ChIP Datasets

Functional Analysis of ZBED6

The TRANSFAC database analysis using ExPlain 3.0 on ZBED6 ChIP data from mouse C2C12 cells revealed that ZBED6 target genes have significant role in different biological processes (Fig. 3). The significant processes include developmental processes (*e.g.* nerves system development, embryonic development, organ development and anatomical structure development), gene expression, regulation of gene expression and metabolic processes etc. Highly enriched (maximum overlapping reads for ZBED6 binding sites) genes showed the role of target genes as developmental proteins (63 out of 756 genes). It also shows that most of the genes that have ZBED6 binding sites downstream to transcription start site TSS are involved in transcriptional regulation (151genes out of 1461 genes) and some are development proteins (87 out of 756). Whereas the genes with ZBED6 binding sites upstream to TSS are also activators (39 genes out of 476 genes).



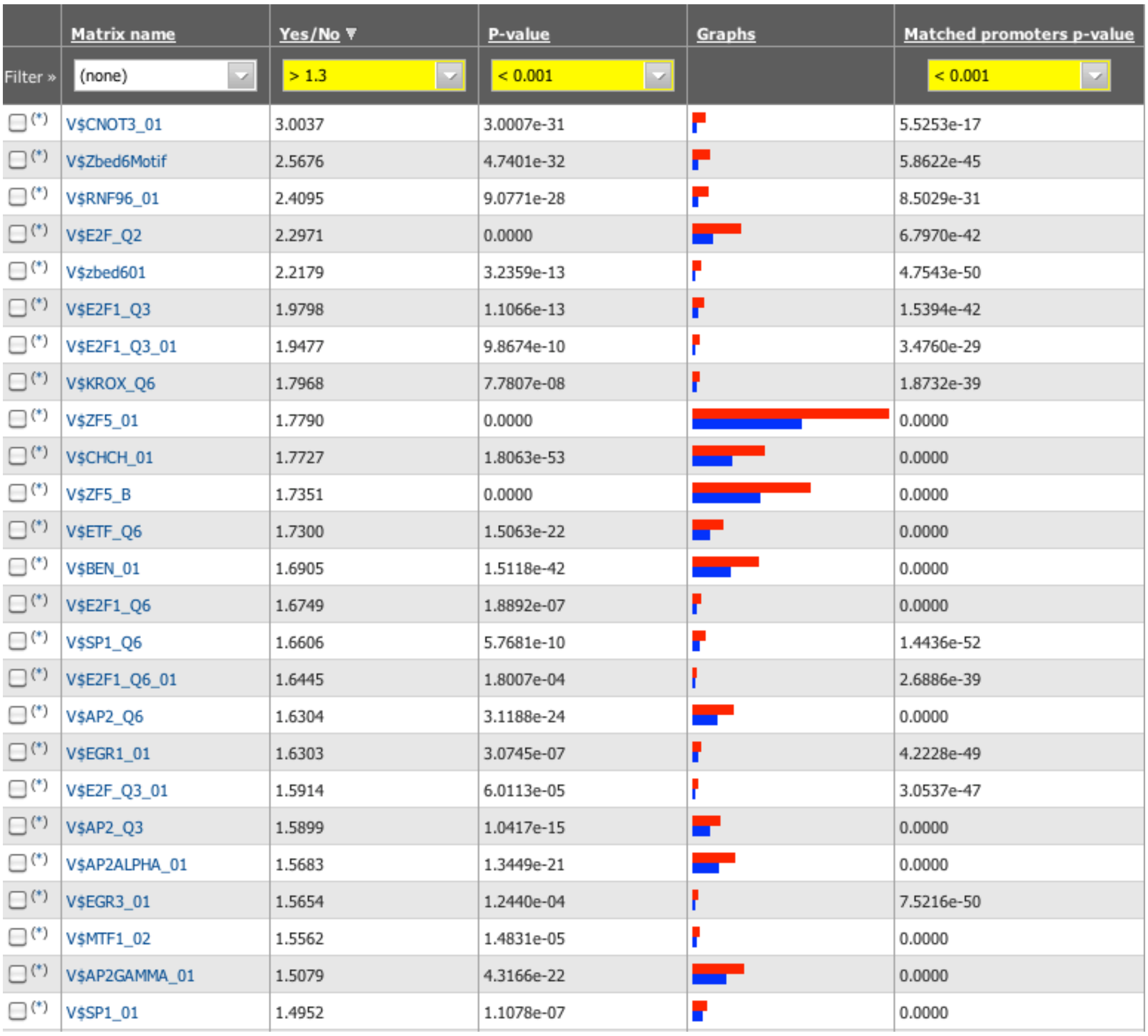
GO Identifier	Gene symbol	Group species	GO Term	Ontology	#Hits in group	Group size	#Hits expected	p-value
GO:0048856	4921517B04Rik, Abr, Acvr2a, Acvr2b, Acz, Ada, Adamts2, Adcy1, Adrb1, Adrbk1, ...	Mouse	anatomical structure development	Biological process	305	4962	208	2.50835e-14
GO:0048731	4921517B04Rik, Abr, Acvr2a, Acvr2b, Acz, Ada, Adamts2, Adcy1, Adrb1, Adrbk1, ...	Mouse	system development	Biological process	290	4640	195	4.94573e-14
GO:0048513	Abr, Acvr2a, Acvr2b, Ada, Adamts2, Adcy1, Adrb1, Adrbk1, Aff1, Aff3, ...	Mouse	organ development	Biological process	236	3498	147	5.50096e-13
GO:0032502	4921517B04Rik, Ablim3, Abr, Acvr2a, Acvr2b, Acz, Ada, Adamts2, Adamts8, Adcy1, ...	Mouse	developmental process	Biological process	380	6805	286	1.0351e-12
GO:0007275	4921517B04Rik, Ablim3, Abr, Acvr2a, Acvr2b, Acz, Ada, Adamts2, Adcy1, Adrb1, ...	Mouse	multicellular organismal development	Biological process	327	5597	235	1.17157e-12
GO:0007417	Abr, Ada, Adcy1, Adrb1, Aff1, Ank2, Ascl1, Atxn1, Cadm1, Cdkn2c, ...	Mouse	central nervous system development	Biological process	86	845	36	4.34754e-12
GO:0007399	4921517B04Rik, Abr, Acz, Ada, Adcy1, Adrb1, Aff1, Ank2, Arhgef7, Ascl1, ...	Mouse	nervous system development	Biological process	158	2165	91	1.43485e-10
GO:0007420	Abr, Ada, Adcy1, Adrb1, Aff1, Ascl1, Atxn1, Cadm1, Cdkn2c, Cited2, ...	Mouse	brain development	Biological process	62	553	24	2.29716e-10
GO:0009790	Abr, Acvr2a, Ada, Ahctf1, Arid5b, Blmh, Bmpr2, Bnc2, Btbd9, Bub1, ...	Mouse	embryonic development	Biological process	122	1532	65	2.36942e-10
GO:0051239	4921517B04Rik, Accn2, Acvr2b, Adra2c, Adrb1, Adrbk1, Ank2, Ankrd2, Arc, Arhgap22, ...	Mouse	regulation of multicellular organismal process	Biological process	148	2025	85	4.36234e-10
GO:0010467	2610305D13Rik, 5930405J04Rik, Ada, Adamts2, Adrbk1, Aff1, Aff4, Ahctf1, Ank2, Ankrd2, ...	Mouse	gene expression	Biological process	261	4428	186	7.15824e-10
GO:0007423	Abr, Aff1, Ascl1, Bcl2l11, Cdh23, Chd7, Cited2, Col2a1, Ece1, Efna5, ...	Mouse	sensory organ development	Biological process	46	355	15	8.66467e-10
GO:0050793	4921517B04Rik, Acvr2b, Ada, Adk, Adrb1, Aff1, Ankrd2, Apbb2, Arhgap22, Arhgef7, ...	Mouse	regulation of developmental process	Biological process	180	2683	113	1.4537e-09
GO:0001654	Aff1, Bcl2l11, Chd7, Cited2, Col2a1, Efna5, Elovl4, Ephb1, Eya3, Foxd1, ...	Mouse	eye development	Biological process	36	251	11	7.38273e-09
GO:0010468	2610305D13Rik, Ada, Adrbk1, Aff1, Aff4, Ahctf1, Ank2, Ankrd2, Arid5b, Arih2, ...	Mouse	regulation of gene expression	Biological process	214	3447	145	9.25275e-09
GO:0001501	Aff1, Atxn1, Bmpr2, Chrdl2, Col2a1, Dlx3, Ece1, Efnb1, Egr2, Fbn1, ...	Mouse	skeletal system development	Biological process	66	680	29	9.53222e-09
GO:0009887	Abr, Acvr2a, Acvr2b, Adamts2, Aggf1, Arhgap22, Ascl1, Bai1, Bcl2l11, Bmpr2, ...	Mouse	organ morphogenesis	Biological process	102	1279	54	1.00382e-08
GO:0043170	2610305D13Rik, 5930405J04Rik, Acacb, Accn2, Acvr2b, Ada, Adam15, Adamts2, Adamts8, Adcy1, ...	Mouse	macromolecule metabolic process	Biological process	367	6941	291	1.16871e-08
GO:0019222	2610305D13Rik, Acacb, Accn2, Acvr2a, Acvr2b, Ada, Adcy1, Adcy8, Adk, Adra2c, ...	Mouse	regulation of metabolic process	Biological process	253	4328	182	3.70622e-08
GO:0060255	2610305D13Rik, Accn2, Ada, Adk, Adrbk1, Aff1, Aff4, Ahctf1, Ank2, Ankrd2, ...	Mouse	regulation of macromolecule metabolic process	Biological process	230	3842	161	3.82211e-08

Figure 3. Role of ZBED6 target genes in Biological processes

Transcription factors in Promoter region of ZBED6 target gene

MatchTM analysis was done on different sets of ZBED6 targets genes (see Materials and Methods ChIP data) in order to find the transcription factor binding sites in the promoter region of ZBED6 target genes. The MatchTM results (transcription factors) were further filtered out using the

Yes/No >1.3 (Yes is the ZBED6 target gene set and No is the control set, here we used mouse house keeping genes as control) and p value < 0.001. The matrices for TFBSs that we obtained from MatchTM are shown in Figure 4.



	Matrix name	Yes/No ▼	P-value	Graphs	Matched promoters p-value
Filter »	(none)	> 1.3	< 0.001		< 0.001
□ (*)	V$CNOT3_01	3.0037	3.0007e-31		5.5253e-17
□ (*)	V$Zbed6Motif	2.5676	4.7401e-32		5.8622e-45
□ (*)	V$RNF96_01	2.4095	9.0771e-28		8.5029e-31
□ (*)	V$E2F_Q2	2.2971	0.0000		6.7970e-42
□ (*)	V$zbed601	2.2179	3.2359e-13		4.7543e-50
□ (*)	V$E2F1_Q3	1.9798	1.1066e-13		1.5394e-42
□ (*)	V$E2F1_Q3_01	1.9477	9.8674e-10		3.4760e-29
□ (*)	V$KROX_Q6	1.7968	7.7807e-08		1.8732e-39
□ (*)	V$ZF5_01	1.7790	0.0000		0.0000
□ (*)	V$CHCH_01	1.7727	1.8063e-53		0.0000
□ (*)	V$ZF5_B	1.7351	0.0000		0.0000
□ (*)	V$ETF_Q6	1.7300	1.5063e-22		0.0000
□ (*)	V$BEN_01	1.6905	1.5118e-42		0.0000
□ (*)	V$E2F1_Q6	1.6749	1.8892e-07		0.0000
□ (*)	V$SP1_Q6	1.6606	5.7681e-10		1.4436e-52
□ (*)	V$E2F1_Q6_01	1.6445	1.8007e-04		2.6886e-39
□ (*)	V$AP2_Q6	1.6304	3.1188e-24		0.0000
□ (*)	V$EGR1_01	1.6303	3.0745e-07		4.2228e-49
□ (*)	V$E2F_Q3_01	1.5914	6.0113e-05		3.0537e-47
□ (*)	V$AP2_Q3	1.5899	1.0417e-15		0.0000
□ (*)	V$AP2ALPHA_01	1.5683	1.3449e-21		0.0000
□ (*)	V$EGR3_01	1.5654	1.2440e-04		7.5216e-50
□ (*)	V$MTF1_02	1.5562	1.4831e-05		0.0000
□ (*)	V$AP2GAMMA_01	1.5079	4.3166e-22		0.0000
□ (*)	V$SP1_01	1.4952	1.1078e-07		0.0000

Figure 4. Matrices of transcription factors in ZBED6 target genes promoter. The transcription factors binding sites matrices using the MatchTM on ZBED6 target gene set. It gives the name of matrix, Yes/No where yes is the target gene set used and No is the mouse housekeeping gene set. The p- value shows the significant overrepresentation sites for the matrices found. The matched promoter p-value gives the significance for the binding sites found in the target gene set promoter. The filters used are yes/no >1.3, p-value <0.001 and matched promoter p-value <0.01. The table is sorted on basis of yes/no in descending order. The profile used was ZBED6h_0.01. The promoter window used is between -1000 to +1000 of TSS.

Using different sets of ZBED6 target genes (Materials and Methods) we obtained the sets of transcription factors binding sites in promoter regions as shown in Table 2 and Figure 5 (Table S1). Using TRANSFAC database and MatchTM, putative common TFs are predicted in the majority of the genes targeted by ZBED6 (Fig. 6). These common TFs include AP2alpha, AP2gamma, ZF5, ZBED6, SP1, E2F1, EGR, CNOT3, ETF, HIC1, RNF96, CHCH and KROX. On the other hand there are few factors that are specific to target gene sets. For example, MTF1 was seen in downstream ZBED6 target genes only while OLF1, NRSE, NFKappaB, LRF and

NGFIC were only seen in upstream targets of ZBED6. The gene set that was far away from TSS (i.e. 500 kb to TSS) showed that they have OCT1, NFAT, IRF1, DMRT2, and NANOG and even did not show some of the common factors (Fig. 6).

Table 2. The Yes/No (ratio of ZBED6 target genes to background genes) values of transcription factors binding sites in the faraway, downstream, and upstream, top1000 and complete Hitlist of ZBED6 target gene sets. These gene sets are named with respect to the distance of ZBED6 binding site to TSS.

Factors	Farawway (yes/no)	Downstream (yes/no)	Upstream (yes/no)	Top1000 (yes/no)	Hitlist (yes/no)
ZF5	2.1389	5.0512	3.6087	5.3733	3.8884
ZBED6	1.6027	3.5419	3.6995	2.0515	1.8628
SP1	1.3354	2.2526	2.3176	1.6971	1.4701
RNF96	3.6875	3.6226	4.7766	3.9323	2.9343
KROX	0	4.2754	3.6314	3.9421	2.1739
HIC1	0	2.1722	2.5575	2.8333	2.0274
ETF	0	2.8075	2.798	2.6323	2.1904
EGR	0	2.2793	1.9869	2.1167	1.4387
E2F1	1.7119	2.9761	2.6375	2.69	2.21
CNOT3	3.5429	3.2603	3.5017	2.8877	2.4815
CHCH	0	2.155	2.1517	2.0727	1.5749
AP2alpha	0	2.7254	2.4691	2.4	2.0067
AP2gamma	0	2.6956	3.2054	4.17	2.0254
OLF1	0	0	2.6725	0	1.8415
NRSE	0	0	2.2489	2.0268	1.7508
NGFIC	0	0	1.9065	0	1.6571
NFKappaB	0	0	1.9996	0	0
LRF	0	0	2.1625	2.0649	1.6768
OCT1	1.4293	0	0	0	0
NANOG	1.4	0	0	0	0
NFAT	1.4843	0	0	0	0
IRF	1.3265	0	0	0	0
HNF1	1.4283	0	0	0	0
HMGIY	1.4581	0	0	0	0
FREAC2	1.78	0	0	0	0
DMRT2	3.5851	0	0	0	0
CDXA	1.7724	0	0	0	0
BRN2	1.8617	0	0	0	0
MTF1	0	2.4825	0	2.183	1.5812

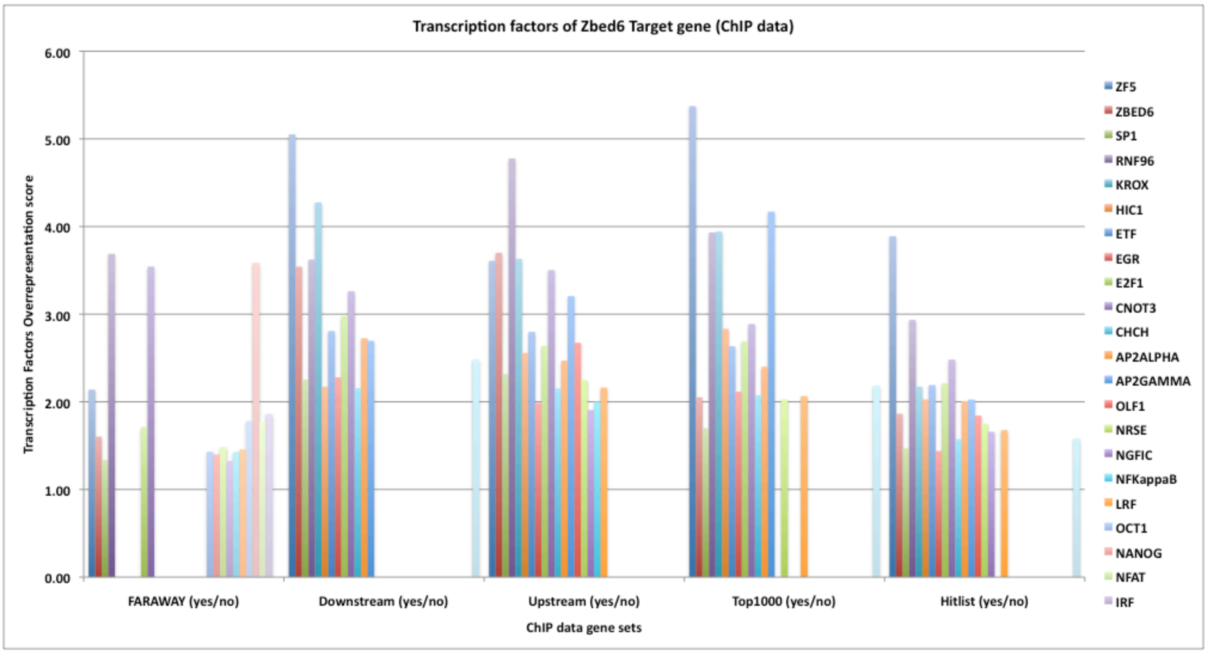


Figure 5. Transcription factors of ZBED6 target genes graphical view. These are the factors that we get in different gene sets of ZBED6 using their yes to no values (Table 2).

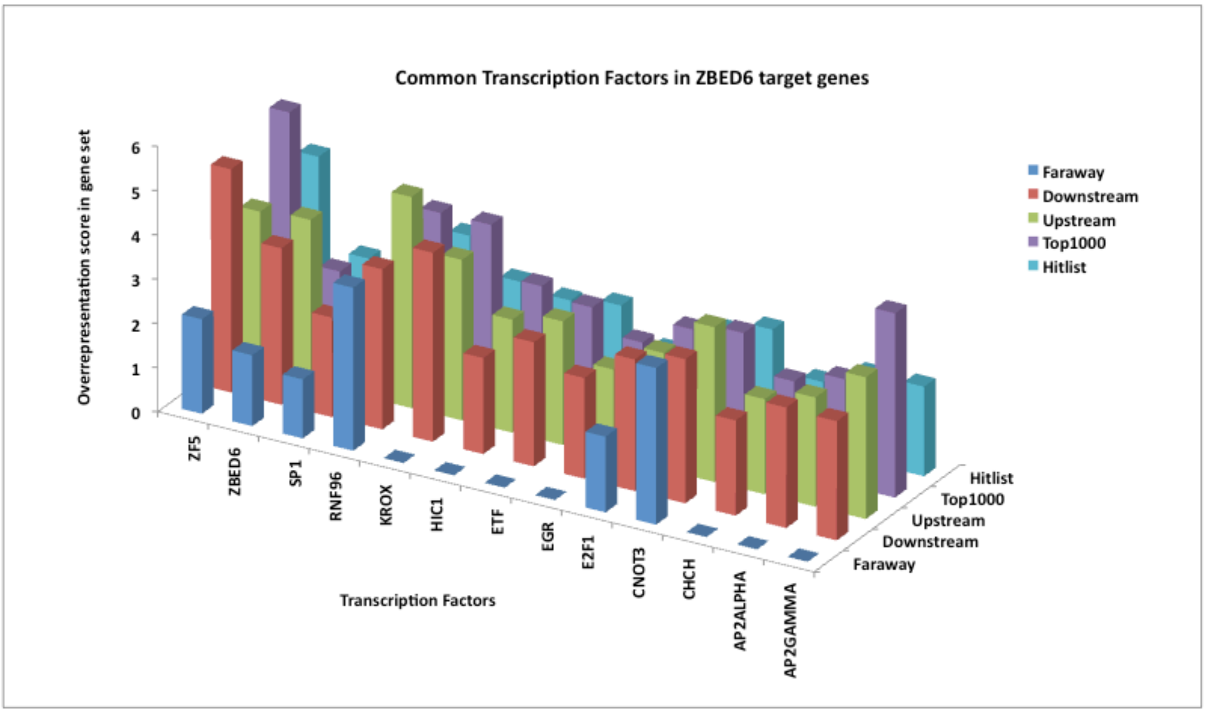


Figure 6. Common transcription factors in the gene set of ZBED6 targets. The figure shows that AP2alpha, AP2gamma, ZF5, ZBED6, SP1, E2F1, EGR, CNOT3, ETF, HIC1, RNF96, CHCH and KROX are the factors that have binding sites in the ZBED6-target genes.

The TRANSFAC results suggest that most of these factors binds near the transcription start site as shown in the Figure 7 that clearly identifies the peak region for the binding of these factors to be close to TSS. This might suggest their role in the regulation of gene and it also supports the previous result (ChIP sequencing results from Markljung et al., 2009) where most of the binding sites for ZBED6 lies close to TSS.

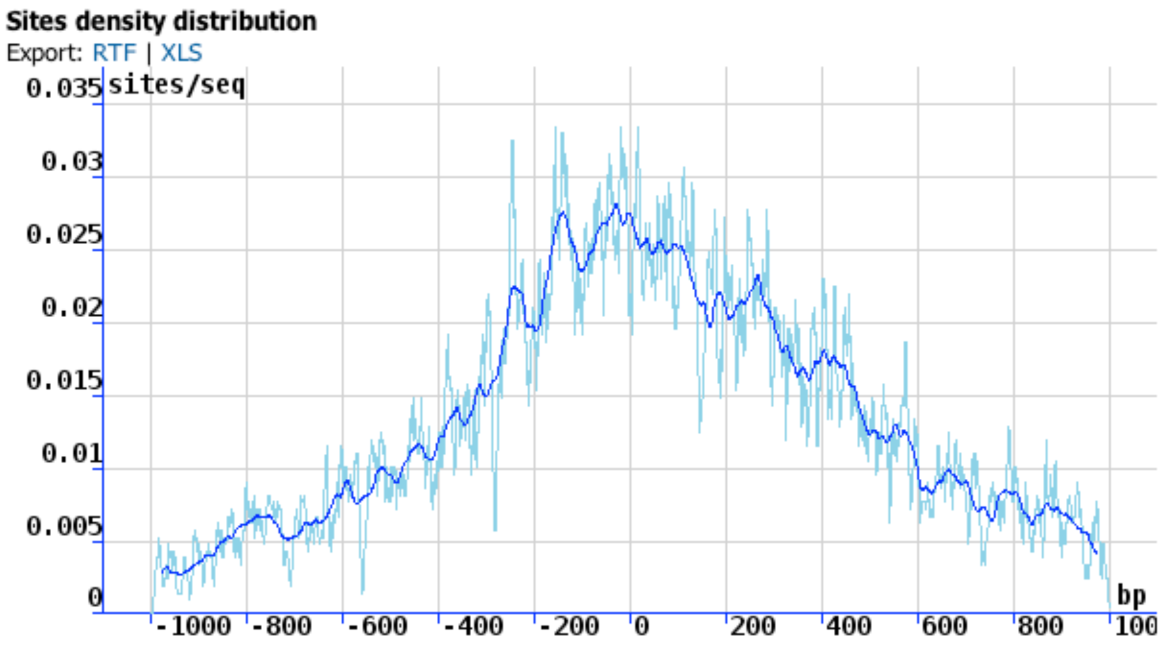
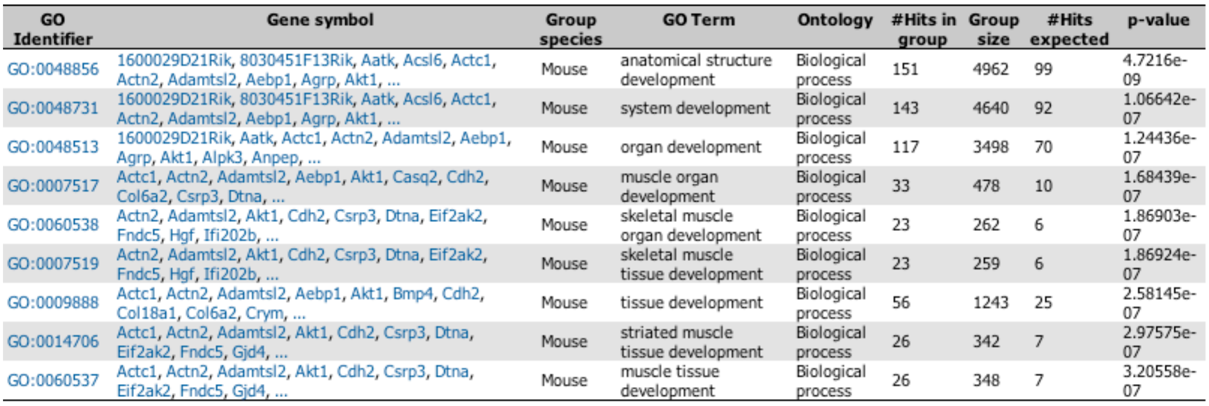


Figure 7. Transcription factor binding sites more in the vicinity of TSS. The horizontal axis shows the promoter region with zero as the TSS. The vertical axis is the binding sites per sequence.

Microarray Datasets

Putative Transcription Factors in ZBED6-silenced Microarray data

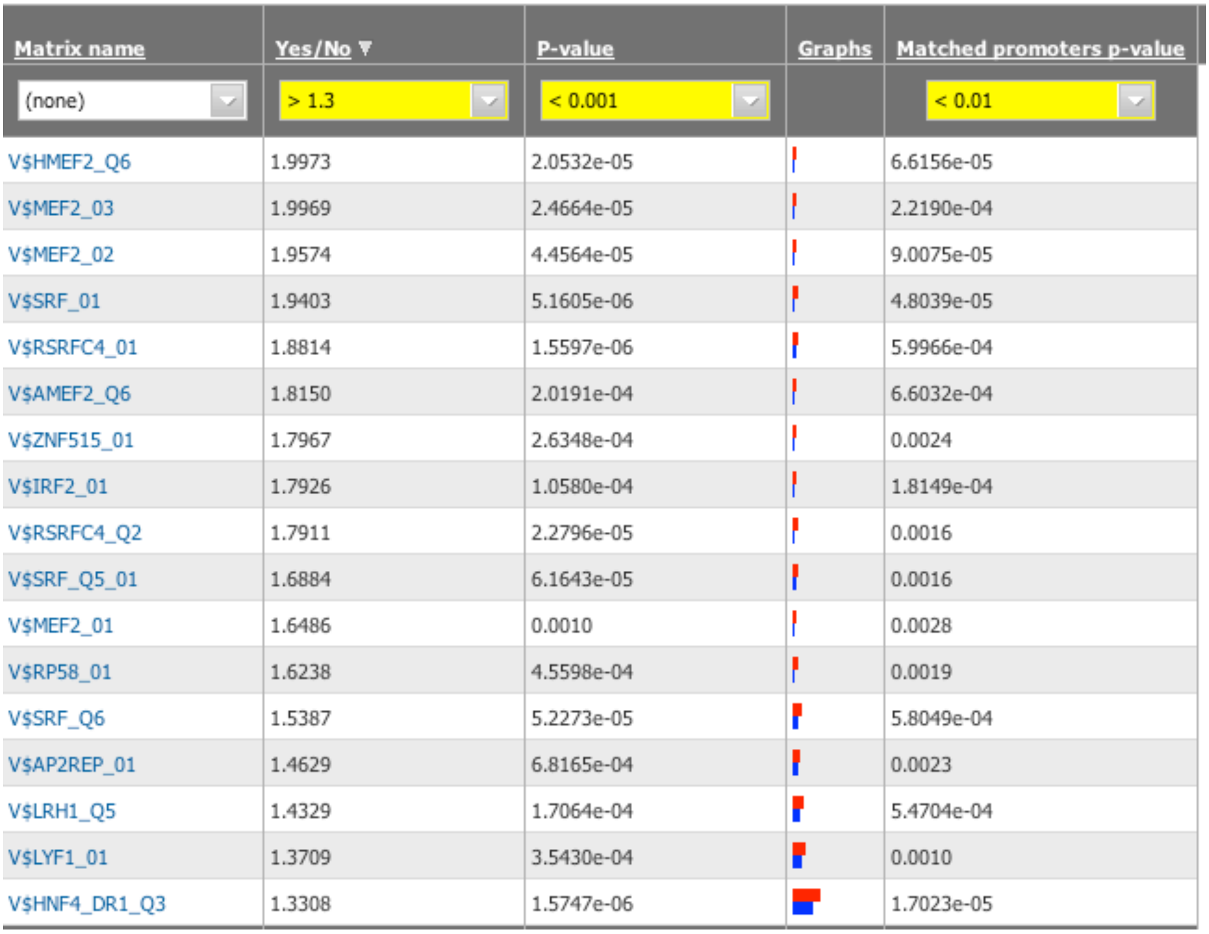
The TRANSFAC database analysis using ExPlain 3.0 on Microarray data from mouse C2C12 cells after silencing of ZBED6 revealed that differentially expressed genes have significant role in different biological processes (Fig. 8). The developmental process is the most significant process including skeletal and striated muscle development, tissue, organ and system development.



GO Identifier	Gene symbol	Group species	GO Term	Ontology	#Hits in group	Group size	#Hits expected	p-value
GO:0048856	1600029D21Rik, 8030451F13Rik, Aatk, Acsl6, Actc1, Actn2, Adamtsl2, Aebp1, Agrp, Akt1, ...	Mouse	anatomical structure development	Biological process	151	4962	99	4.7216e-09
GO:0048731	1600029D21Rik, 8030451F13Rik, Aatk, Acsl6, Actc1, Actn2, Adamtsl2, Aebp1, Agrp, Akt1, ...	Mouse	system development	Biological process	143	4640	92	1.06642e-07
GO:0048513	1600029D21Rik, Aatk, Actc1, Actn2, Adamtsl2, Aebp1, Agrp, Akt1, Alpk3, Anpep, ...	Mouse	organ development	Biological process	117	3498	70	1.24436e-07
GO:0007517	Actc1, Actn2, Adamtsl2, Aebp1, Akt1, Casq2, Cdh2, Col6a2, Csrp3, Dtna, ...	Mouse	muscle organ development	Biological process	33	478	10	1.68439e-07
GO:0060538	Actn2, Adamtsl2, Akt1, Cdh2, Csrp3, Dtna, Eif2ak2, Fndc5, Hgf, Ifi202b, ...	Mouse	skeletal muscle organ development	Biological process	23	262	6	1.86903e-07
GO:0007519	Actn2, Adamtsl2, Akt1, Cdh2, Csrp3, Dtna, Eif2ak2, Fndc5, Hgf, Ifi202b, ...	Mouse	skeletal muscle tissue development	Biological process	23	259	6	1.86924e-07
GO:0009888	Actc1, Actn2, Adamtsl2, Aebp1, Akt1, Bmp4, Cdh2, Col18a1, Col6a2, Crym, ...	Mouse	tissue development	Biological process	56	1243	25	2.58145e-07
GO:0014706	Actc1, Actn2, Adamtsl2, Akt1, Cdh2, Csrp3, Dtna, Eif2ak2, Fndc5, Gjd4, ...	Mouse	striated muscle tissue development	Biological process	26	342	7	2.97575e-07
GO:0060537	Actc1, Actn2, Adamtsl2, Akt1, Cdh2, Csrp3, Dtna, Eif2ak2, Fndc5, Gjd4, ...	Mouse	muscle tissue development	Biological process	26	348	7	3.20558e-07

Figure 8. Role of ZBED6-silenced genes in Biological processes

The results from MatchTM on microarray data suggest that SRF and MEF2 are most significantly expressed transcription factors in ZBED6-silenced C2C12 cells as shown in Figure 9.



Matrix name	Yes/No ▼	P-value	Graphs	Matched promoters p-value
(none)	> 1.3	< 0.001		< 0.01
V$HMEF2_Q6	1.9973	2.0532e-05		6.6156e-05
V$MEF2_03	1.9969	2.4664e-05		2.2190e-04
V$MEF2_02	1.9574	4.4564e-05		9.0075e-05
V$SRF_01	1.9403	5.1605e-06		4.8039e-05
V$RSRFC4_01	1.8814	1.5597e-06		5.9966e-04
V$AMEF2_Q6	1.8150	2.0191e-04		6.6032e-04
V$ZNF515_01	1.7967	2.6348e-04		0.0024
V$IRF2_01	1.7926	1.0580e-04		1.8149e-04
V$RSRFC4_Q2	1.7911	2.2796e-05		0.0016
V$SRF_Q5_01	1.6884	6.1643e-05		0.0016
V$MEF2_01	1.6486	0.0010		0.0028
V$RP58_01	1.6238	4.5598e-04		0.0019
V$SRF_Q6	1.5387	5.2273e-05		5.8049e-04
V$AP2REP_01	1.4629	6.8165e-04		0.0023
V$LRH1_Q5	1.4329	1.7064e-04		5.4704e-04
V$LYF1_01	1.3709	3.5430e-04		0.0010
V$HNF4_DR1_Q3	1.3308	1.5747e-06		1.7023e-05

Figure 9. Matrices of transcription factors in differentially expressed gene sets of ZBED6 silenced Microarray data. The promoter window used is between -500 to +100 of TSS. The Yes is the set of differentially expressed genes while No is the non-differentially expressed genes used as control set. The filters are applied to get the most significant matrices.

The transcription factors that we obtained from TRANSFAC database MatchTM analysis shows that they are functionally important as shown in Table 3.

Table 3. Transcription factors regulating ZBED6 target genes (selected factors). The table shows the factors with their functions and reference.

Transcription Factor Name	Function	Reference
AP2 Alpha (Activating protein alpha)	Inhibits the growth of cells by inducing cell cycle arrest and apoptosis and crucial role in tumorigenesis	Orso et al., 2008
ZF5	Transcriptional repressor in mouse c myc promoter	Obata et al., 1999
ZBED6	Repressor of IGF2 gene expression	Markljung et al., 2009
SP1 (Sequence specific transcription factor 1)	Required for expression of variety of genes involved in cell proliferation, apoptosis, development and differentiation.	Kaczynski et al., 2003

E2F1	Crutial role in control of cell cycle and role as tumor suppressor protein.	Tao et al., 1997
OCT1	Octamer binding transcription factor (POU domain class 2 transcription factor)	Marecki et al., 2001
NFAT	Nuclear factor of activated T cells mediated signal transduction pathways.Important in skeletal, cardiac muscle and nervous systems.	Macian, 2005
IRF1 (Interferon regulatory factor 1)	Activates transcription of interferon alpha and beta. IRF1 plays role in regulating apoptosis and tumor suppression.	Marecki el al., 2001
HIC1 (hypermethylated in cancer 1)	Gene encoding the zinc finger transcription factor belonging from the POZ family. It is also candidate tumor suppressor gene. Attenuates Wnt signaling	Valenta et al., 2006

Long Non-Coding RNA

To find the long non-coding RNA that may be regulated by the ZBED6 we first find the distance of ZBED6 peak to the predicted long non-coding RNAs. In this case we have taken ZBED6-binding fragments with 500 kb distance from the TSS. Out of 2499 binding fragments of ZBED6 we obtained a set of 300 regions (Table S2).

For the long non-coding RNAs we used two datasets Ponjavic dataset generated by FANTOM consortium and Guttman for the intergenic non-coding RNAs were further filtered out to eliminate the all protein coding genes (Carninci et al., 2005; Ponjavic et al., 2007; Guttman et al., 2009). The obtained datasets Guttman and Ponjavic have 1408 and 2168 lncRNAs respectively (De Santa et al., 2010). Using the perl script (see Materials and Methods) we find the ZBED6 peaks in the vicinity of lncRNAs TSS as shown in Table 4.

Table 4. Shows the number of ZBED6 peaks with in the Range of long non-coding RNA Transcription start site.

Range	Guttman (1408)	Ponjavic (2168)
2kb	5	0
5kb	8	0
10kb	12	0
20kb	13	1
30kb	14	5
40kb	14	6
50kb	18	12
60kb	23	14
70kb	23	17
80kb	27	21
90kb	29	30
100kb	35	39

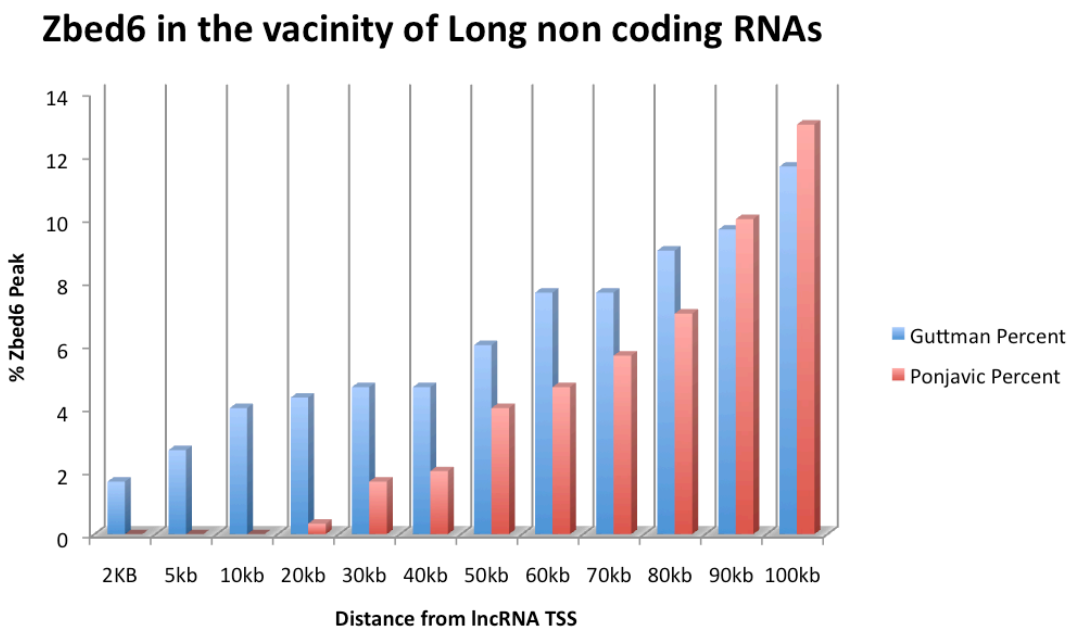


Figure.10 Percentage of ZBED6 peaks near long non-coding RNAs TSS (Guttman and Ponjavic datasets).

The graph (Fig. 10) shows that only about 2% of ZBED6 binding sites lies within 2 kb region of lncRNAs TSS. As the distance from the TSS increases the ZBED6 binding sites also increases (Table S3). This tells us that ZBED6 might not be present in the promoter region of non-coding RNAs but it could be present in the enhancer region of these non-coding RNA genes. Since the presence of ZBED6 at these regions could not be correlated with the protein coding genes, we can assume that maybe ZBED6 is playing the role of enhancer or silencer elements.

Table 5. ZBED6 sites within 30 kb of lncRNAs.

ZBED6 Peak	Chromosome	Start_ncRNA	End_ncRNA	ZBED6 Distance_TSS (30 kb)	ZBED6 Enrichment*	mRNA_Genbank
151643270	chr5	151621840	151622708	21430	26	AK029411
48959270	chr14	48980247	48980948	-20977	24	AK007311
143954645	chr3	143983177	143984712	-28532	20	AK035087
143286560	chr3	143313273	143315051	-26713	16	AK034713
95129690	chr10	95143442	95143758	-13752	16	AK009289

* Enrichment is the overlapping reads obtained from ZBED6 ChIP data (Markljung et al., 2009)

The mRNA from Gene bank (Table 5) for the non-coding RNAs, those have the ZBED6 peak present between 10 kb to 30 kb region. The genome view of one of the mRNA AK029411 is shown in Figure 11 A and B that is obtained by using UCSC and Ensemble browsers respectively.

Figure 11 B also shows presence of snoRNA genes like SNORA17 and few U3 (found predominantly in nucleolus), U4, U5 and U6 present in that region. These ncRNAs are seen in vicinity of few other mRNA also.

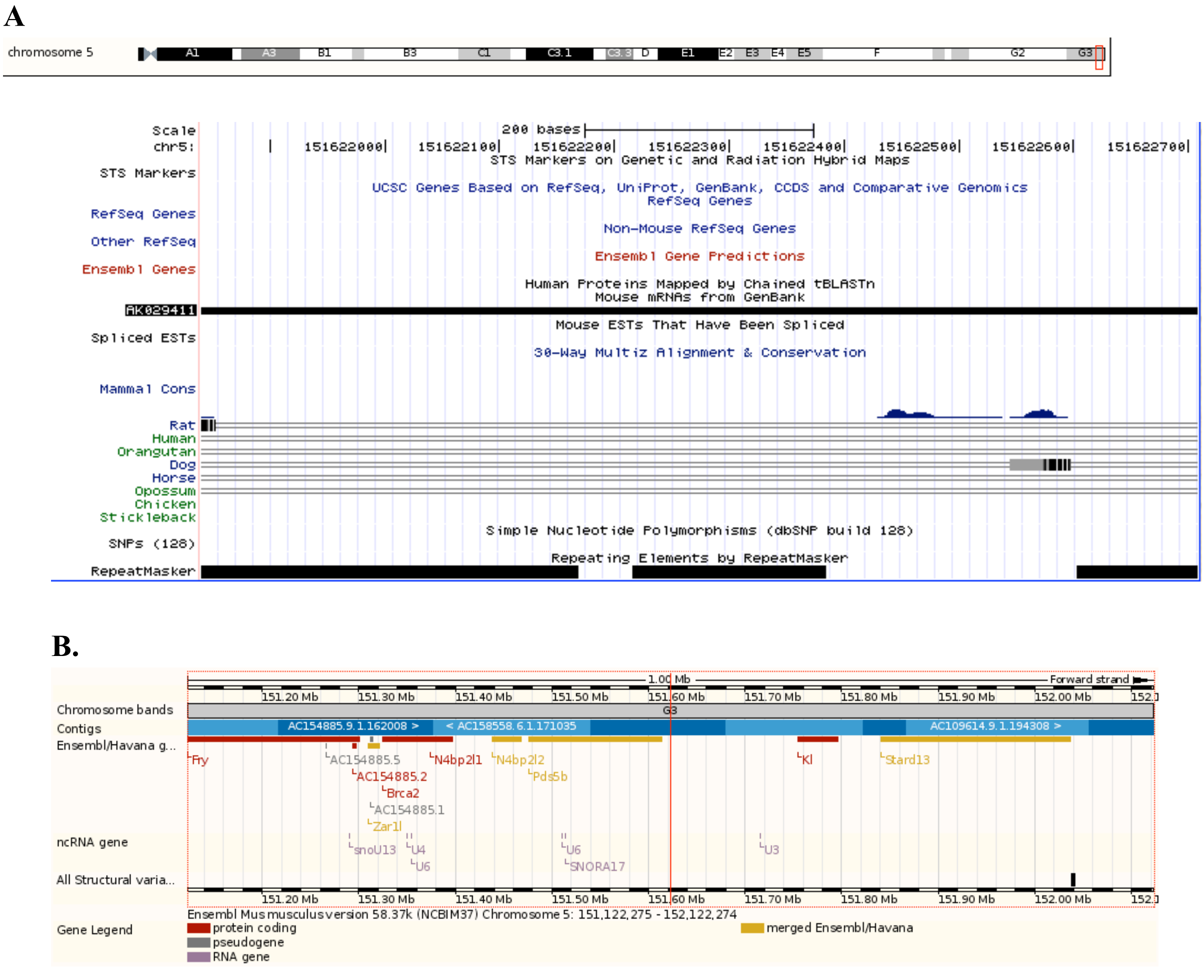


Figure 11. A. Long non coding mRNA AK029411 present at distance of 30 kb from ZBED6 peak using UCSC genome browser (http://genome.ucsc.edu/). B. The region around mRNA AK029411 shows some small nucleolar organizing RNAs snoRA1 , U3, U4, U5 and U6 using Ensembl genome browser (http://www.ensembl.org/Mus_musculus/Info/Index).

DISCUSSION

ChIP experiment

The Chromatin Immuno precipitation ChIP experiment that was carried out in mouse C2C12 cells showed 2499 binding fragments for ZBED6. The experiment was carried out using the AB SOLiD technology (Markljung et al., 2009). ZBED6 binding fragments from the ChIP data were used as the input set for bioinformatics analysis using the TRANSFAC database and its MatchTM tool (professional version). The purpose of these bioinformatics analyses was to find the other TFs that might be co regulating the expression of genes with a novel transcription factor ZBED6. ZBED6 has shown to be involved in many biological processes including in developmental processes (e.g. nerves system development, embryonic development, organ development, anatomical structure development etc.), gene expression, regulation of gene expression and metabolic processes (Fig. 3). Most of the genes that are downstream target of ZBED6 have role in transcriptional regulation. Where as few genes that are upstream target of ZBED6 show that they are activator proteins. This suggests the role of ZBED6 in transcription regulation and its involvement in many important developmental processes in mouse C2C12 cells.

As the experiment was done in the mouse myoblast C2C12 cells we do not currently know much about its role in other cell types. However, ZBED6 is expressed in most cell types and is likely to have many more target genes. The CpG Island in *IGF2* intron 3 was found to be non-methylated at a region centered at the QTN in case of skeletal muscle while in case of liver it was methylated (Van Laere at al., 2003). The epigenetic state of chromatin is therefore also very critical in the role of ZBED6.

TRANSFAC: MatchTM

In order to find the transcription factors binding sites TFBS in the ZBED6 targeted genes, we have used MatchTM tool that uses the position weight matrices provided by TRANSFAC library. A number of TFBSs were seen to be overrepresented in the promoter regions of ZBED6 target genes (Fig. 4). We have used promoter window of -1000 upstream to TSS and +1000 downstream to the TSS. A number of filters (p value < 0.001, yes/no >1.3) were also used to find the most significant transcription factors.
The results of MatchTM can be biassed based on a number of reasons:

- The gene sets (Material and Methods, ChIP datasets) that we have selected from the ZBED6 target list might have missed out some important genes.
- The background or control set mouse house keeping genes that are provided by TRANSFAC database may have some genes missing.
- There is a possibility that we might have missed out some important factors because of the filters that we have applied.
- The gene expression is very complex process and it may or may not be same in different cell types. So the predicted transcription factors can be different or same in other scenario.
- The transcription factors might not always bind to the fixed motif or TFBSs because there are other environmental and cellular components involved in the regulation of gene expression. This might affect the results that we obtain from the bioinformatics analysis.
- Although the results were very much consistent with the previous results of ZBED6 functional analysis and role in important biological processes but we cannot rule out the possibility of biasess due to the use of parameters and cut-off values.

Transcription factors in ZBED6 target genes promoters

The significant common transcription factors from the MatchTM analysis on ChIP data of ZBED6 in mouse C2C12 cells are AP2alpha, AP2gamma, ZF5, ZBED6, SP1, E2F1, EGR, CNOT3, ETF, HIC1, RNF96, CHCH and KROX (Fig. 6 and Table 3). The majority of ZBED6-targeted genes show the binding sites for these transcription factors. This might suggest that ZBED6 along with these factors co regulates the expression of number of genes that are involved in many important biological processes. The results also revealed that majority of ZBED6 sites were closer to the ZF5 sites. In some cases the two binding sites were also overlapping.

On the other hand there are few factors that are specific to target gene sets. For example, MTF1 was seen in downstream ZBED6 target genes only while OLF1, NRSE, NFKappaB, LRF and

NGFIC were only seen in upstream targets of ZBED6. The gene set that was far away from TSS (i.e. 500 kb to TSS) showed that they have OCT1, NFAT, IRF1, DMRT2, and NANOG and even did not showed some of the common factors (Fig. 6).

The TRANSFAC results suggests that most of these factors binds near the transcription start site as shown in the Figure 7 that clearly identifies the peak region for the binding of these factors to be close to TSS. This might suggest their role in the regulation of gene. It also supports the previous result (ChIP sequencing results from Markljung et al., 2009) where most of the binding sites for ZBED6 lies close to TSS. The transcription factors of Faraway gene set shows some contradiction to these binding sites in the vicinity of TSS.

One important point to note here is the Faraway gene set (Materials and Methods, ChIP dataset) that have the ZBED6 binding sites at a distance of 500 kb way from TSS showed most of the factors that were not present in the other ZBED6 target gene sets (upstream or downstream). Furthermore, a few factors that are common to all sets are not present in this faraway set. One possibility is that these regions could be the enhancer regions for most of the genes present in the target list of ZBED6 and these factors may be enhancer elements. The other possibility is that these faraway regions are promoter regions of non-coding RNAs that are largely present in the intergenic regions of mammalian genomes (Birney et al., 2007; Prasanth et al., 2007). The hypothesis was further tested using the noncoding RNAs datasets and ZBED6 faraway regions.

Long non-coding RNA

The ChIP sequencing experiment using the C2C12 myoblast revealed more than 1200 of the ZBED6 binding sites that occurred within 5 kb of the TSS of an annotated gene (Markljung et al., 2009). Using the gene ontology analysis it was revealed that these genes are associated with important biological processes including development, transcriptional regulation and cell differentiation. But those regions (300 out of 2499) that are at a distance of more than 500 kb from the known protein coding genes are still to be analyzed for their functional significance. The ZBED6 binding sites at this large distance from known TSS suggest that ZBED6 might have a regulatory role acting as a TF at enhancer or at distal promoter elements.

The results of non-coding RNA (Fig. 10) shows that only 2% of ZBED6 sites in faraway regions lies within 2 kb of lncRNAs TSS. As the distance from the TSS of lncRNA increases the number of ZBED6 binding sites also increases. Table 5 shows some of the selected mRNA of these predicted non-coding RNAs having the ZBED6 binding sites at a distance of 30 kb away from TSS. Figure 11A and 11B shows that there are also small nuclear organizing RNAs in the vicinity of these lncRNAs. It is possible that ZBED6 has a functional role in regulating transcription of these small nucleolar organizing RNAs like snoRA17 and U3, U4 etc. U3 is a small nucleolar organizing RNA expressed predominantly in the nucleolus. This is an interesting possibility as we know that the localization of ZBED6 is also within nucleolus (Markljung et al., 2009).

Another possibility is that ZBED6 is not actually regulating transcription from promoter regions of these lncRNAs but instead by binding to enhancer regions. The other possibility is that these regions are the enhancer regions of many ZBED6 target genes (protein coding) and it remains to be established whether ZBED6 is functioning as an enhancer factor in those regions.
We also cannot rule out the possibility that maybe it is present there due to the binding to the Histone H3K4me3/H3K36me3 chromatin signatures of characteristic active genes (Guttman et al., 2009).

Further experiments are needed to confirm these results to determine if there is really some functional significance of these intergenic (faraway) regions or whether they are artefactous hits from the ChIP experiments.

Transcription factors in differentially expressed genes in ZBED6-silenced C2C12 cells

In mouse myoblast C2C12 cells, ZBED6 was silenced using siRNA and the microarray analysis was done for day 2, day 4 and both days time point. The genes that showed the differential expression were analysed using TRANSFAC database. The functional analysis (Fig. 8) showed that majority of genes that are potentially regulated by ZBED6 have role in the development of skeletal muscle, tissues and organs. The results (Fig. 9) showed that SRF and MEF2 in all cases were common transcription factors in these gene sets while in some cases Myogenin, MYOD, LYF and MAZ (muscle specific) transcription factors were also seen. SRF and MEF2 have role in muscle development. The factors that were found on day 2 have more role in cell cycle and proliferation of cells. While the factors that were found for both day time points showed that they are more involved in the process of differentiation of cells. This result also supports the role of IGF2 in muscle development (Florini et al., 1991). The role of ZBED6 as repressor was first of all identified in the *IGF2* gene at the QTN region (Van Laere et al., 2003). However, based on available microarray results following ZBED6-silencing it should be stressed that ZBED6 may function not only as a repressor but also as an activator of transcription for a number of genes in C2C12 cells. The functional role of ZBED6 as a transcriptional regulator of other genes besides *IGF2* remains to be established. Furthermore, target genes for ZBED6 in other cells types will be determined and the role of ZBED6 in other cell types except myoblasts needs to be determined.

Our current proposed model for ZBED6 function as repressor or activator is summarized in Figure 12. The IGF2 locus indicates that ZBED6 acts primarily as a repressor but there is a possibility that it acts as activator under certain circumstances.

The repressor model shows the binding sites of the transcription factors AP2, HIC1, MEF2, SP1 and ZBED6 in the promoter region (-500 bp to +500 bp from TSS) of the upregulated genes in ZBED6 silenced C2C12 cells. These transcription factors have crutial role in the cell cycle arrest and tumorigenesis (Table 3). MEF2 which is Myocyte enhancer factor 2 controls the muscle specific and growth factor inducible genes. HIC1 is epigenetically inactivated in cancer and encodes the zinc finger transcription factor belonging to the POZ family. It is also a candidate tumor suppressor gene. ZBED6 along with these transcription factors may repress the expression of number of genes involved in the tumorigenesis. There is a need to confirm this using functional biological experiments as the presence of binding sites may not be the proof of regulation of expression.

The activator model shows the binding sites of the transcription factors E2F1, ZF5, SP1,KROX, OCT1, IRF1 and ZBED6 in the promoter region of downregulated genes under ZBED6 silenced C2C12 cells. It is observed that ZF5 (ZiN POZ domain contain transcription factor) sites were overlapping with the ZBED6 binding sites for the majority of genes and SP1 and KROX binding sites were also in the close proximity. IRF1 and OCT1 were observed together in a number of genes. All these transcription factors have GC rich binding sites and the majority belong to the family of zinc finger transcription factors. There is a possibility that ZBED6 along with other zinc finger transcription factors coregulates the expression of genes involved in important biological processes.

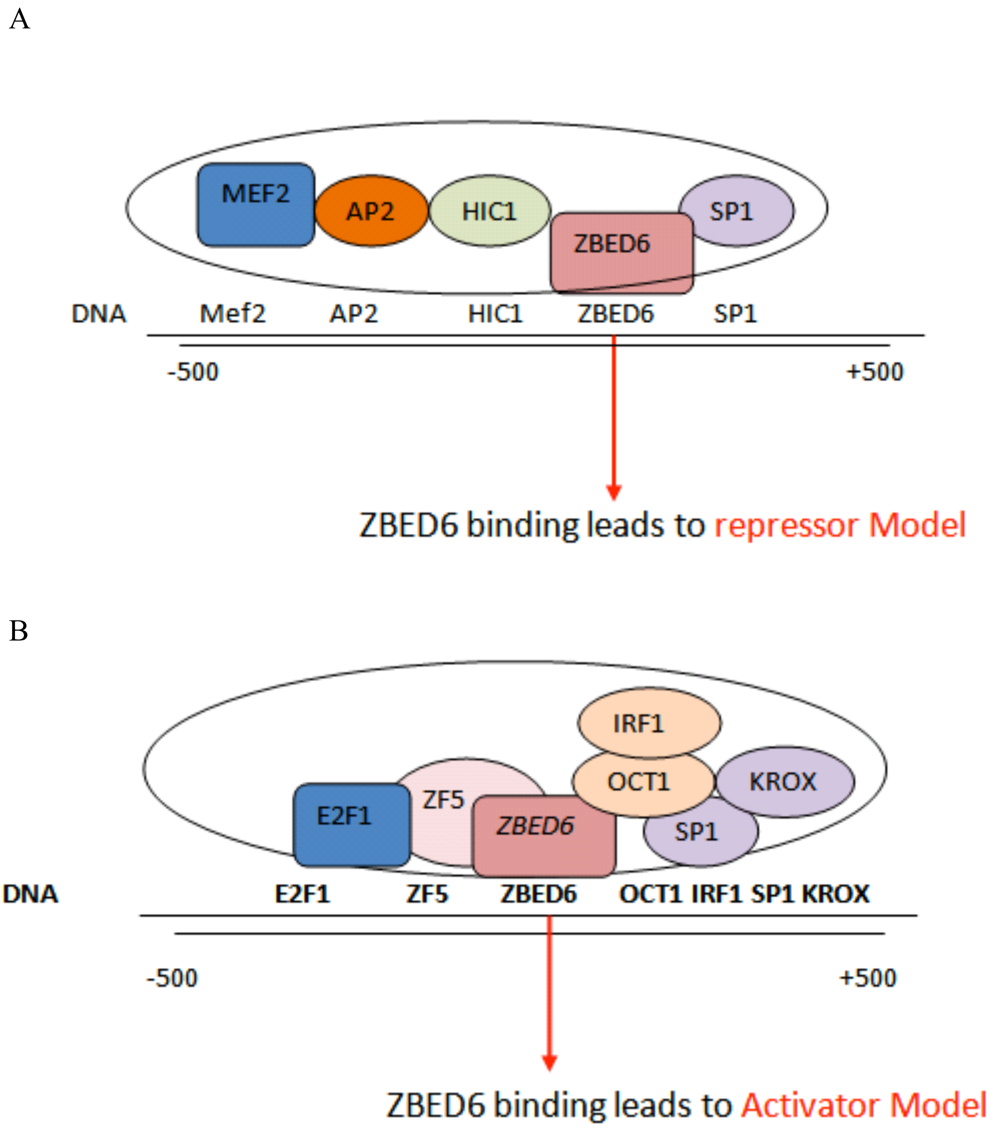


Figure 12 A. Proposed ZBED6 repressor model. TFBS in the promoter region of Upregulated gene set in siRNA silenced ZBED6 C2C12 cells. B. Proposed ZBED6 activator model. TFBS in the promoter region of Down regulated gene set in siRNA silenced ZBED6 C2C12 cells.

CONCLUSION

The ZBED6 factor along with a number of transcription factors (e.g. HIC1, AP2alpha, ZF5, E2F1, KROX, SP1 majority belonging from zinc finger family and having GC rich binding sites) co-regulates the expression of genes that have roles in important biological processes (proliferation, differentiation and development). The majority of factors found in the promoter regions of ZBED6 target genes show that they are involved in tumor suppression. This suggests that ZBED6 might be involved in the suppression of tumorigenesis a notion supported by the fact that many of the identified ZBED6 target genes have established roles as tumor suppressors. In the absence of ZBED6 the transcription factors like SRF and MEF2 are involved in the cell differentiation and muscle development.

ZBED6 also show some enhancer activity with the transcription factors OCT1, IRF1 and NFAT. These factors might together with ZBED6 be co-regulating the expression of number of genes. ZBED6 might be present in the enhancer region of long non-coding RNAs and regulating their expression. These ncRNAs does have roles in the different types of carcinomas. ZBED6 might be regulating their expression as well.

FUTURE PROSPECTS

In future the significance of these transcription factors can be confirmed by looking into the evolutionary conserved ZBED6 target genes. There is also a need to confirm the co-regulation of these factors biologically by performing functional experiments as the bioinformatics analysis only provides a prediction that needs to be experimentally confirmed. For the correlation of long non-coding RNAs with ZBED6 the complete set of lncRNAs including known and unknown should be taken into consideration to conclude the study. This may be possible as more lncRNAs become described as a result of global transcriptome experiments.

ACKNOWLEDGMENTS

I would like to thank the following persons for helping me in one way or another in writing this thesis:

- My supervisors Goran Andersson, Erik Bongcam Ruddloff and Leif Andersson for interesting discussions and opinions on results and written work. I am thankful to them for providing this interesting research opportunity and always making time for questions and discussions to come to this final report.
- Higher Education Commission, Pakistan for providing me the financial support to pursue my studies in Sweden.
- Lin Jiang, Carl Johan Rubin and Alvaro Martinez for guiding me whenever I needed their help.
- All the members of ZBED6 group in their guidance and complete support in this study.
- Special thanks to all my friends for their valuable suggestions.

REFERENCES

Amaral, P.P, Mattick, J.S. (2008) Noncoding RNA in development. *Mamm Genome*, **19:**454-492.

Amaral, P.P., Mattick, J.S. (2008) Noncoding RNA in development. *Mamm Genome*, **19:**454-492.

Amarger, V., Nguyen, M., Van Laere, A.S., Braunschweig, M., Nezer, C., Georges, M., Andersson, L. (2002) Comparative sequence analysis of the Insulin-IGF2–H19 gene cluster in pigs. *Mamm. Genome* **13:** 388–398

Aravind,L.(2000) The BED finger, a novel DNA-binding domain in chromatinboundary-element-binding proteins and transposases. *Trends Biochem Sci* **25:**421–423.

Ashe, H. L., Monks, J., Wijgerde, M., Fraser, P., Proudfoot, N. J. (1997) Intergenic transcription and transinduction of the human beta-globin locus. *Genes Dev,* **11:** 2494–2509

Ashe, H.L., Monks, J., Wijgerde, M., Fraser, P., Proudfoot, N.J. (1997) Intergenic transcription and transinduction of the human beta-globin locus. *Genes Dev* **11:** 2494–2509.

Birney, E, Stamatoyannopoulos, J. A, Dutta, A., Guigo, R., Gingeras, T. R, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799-816.

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:** 1021–1029.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C, et al. (2005) The transcriptional landscape of the mammalian genome. *Science* **309:** 1559-1563.

Chang, L.W., Nagarajan, R., Magee, J.A., Milbrandt, J., Stormo, G.D. (2006) A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profile. *Genome Res.* **16:** 405-413.

Chen, K., Baxter, T., Muir, W. M., Groenen, M. A., Schook, L. B. (2007) Genetic resources, genome mapping and evolutionary genomics of the pig (Sus scrofa). *Int J Biol Sci,* **3:** 153-165.

Coolican, S. A., Samuel, D. S., Ewton, D. Z., McWade, F. J., Florini, J. R. (1997) The mitogenic and myogenic actions of insulin-like growth factors utilize distinct signaling pathways. *J Biol Chem* **272:** 6653-6662.

Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J., Palsson, B.Q. (2004) Integrating high throughput and computational data elucidates bacterial networks. *Nature* **429:** 92-96.

De Santa ,F, Barozzi,I., Mietton,F., Ghisletti,S., Polletti,S., et al. (2010) A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biol* **8(5):** e1000384.

Dieci, G., Fiorino, G., Castelnuovo ,M., Teichmann, M., Pagano, A. (2007) The expanding RNA polymerase III transcriptome. *Trends in Genetics,* **23 (12):** 614–22.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., Pagano, A. (2007) The expanding RNA polymerase III transcriptome. *Trends in Genetics,* **23 (12):** 614–22.

Dinger, M.E, Amaral, P.P, Mercer, T.R., Pang, K.C., Bruce, S.J., Gardiner, B.B., Askarian-Amiri, M.E, Ru, K., Solda, G., Simons, C. et al. (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res*, **18:**1433-1445.

Dinger, M.E., Amaral, P.P, Mercer, T.R., Pang, K.C., Bruce, S.J.,Gardiner, B.B., Askarian-Amiri M.E., Ru K, Soldà G, Simons, C. (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res* **18:**1433-1445.

Eden, S., Constancia, M., Hashimshony, T., Dean, W., Goldstein, B., Johnson, A.C., Keshet, I., Reik, W., Cedar, H. (2001) An upstream repressor element plays a role in Igf2 imprinting. *EMBO J.* **20:**3518–3525.

Engert, J.C., Berglund, E.B., Rosenthal, N. (1996) Proliferation precedes differentiation in IGF-I-stimulated myogenesis. *J Cell Biol* **135:** 431-440.

Florini, J. R., Ewton, D. Z., McWade, F. J.(1995) IGFs muscle growth, and myogenesis. *Diabetes Rev.* **3:** 73–92.

Florini, J. R., Magri, K.A., Ewton, D.Z., James, P.L., Grindstaff, K., Rotwein, P.S. (1991) Spontaneous differentiation of skeletal myoblasts is dependent upon autocrine secretion of insulin-like growth factor-II. *Biol Chem* **266:** 15917-15923.

Fu, X., Ravindranath, L., Tran, N., Petrovics, G., Srivastava, S. (2006) Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1. *DNA and Cell Biology,* **25 (3):** 135–41.

Fu, X., Ravindranath, L., Tran, N., Petrovics, G., Srivastava, S. (2006) Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1. *DNA and Cell Biology,* **25 (3):** 135–41.

Fujii, J., Otsu, K., Zorzato, F., de Leon, S., Khanna, V.K., Weiler, J.E., O'Brien, P.J., MacLennan, D.H. (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science,* **253:** 448-451.

Gerrard, D. E., Okamura, C. S., Ranalletta, M. A.,Grant, A. L. (1998) Developmental expression and location of IGF-I and IGF-II mRNA and protein in skeletal muscle. *J Anim Sci* **76:** 1004-1011.

Giuffra, E., Kijas, J.M.H., Amarger, V., Carlborg, Ö., Jeon, J.T., Andersson, L. (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics,* **154:** 1785-1791.

Greally, J. M., Guinness, M. E., McGrath, J., Zemel, S. (1997) Matrix-attachment regions in the mouse chromosome 7F imprinted domain. *Mamm. Genome* **8:** 805–810.

Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R., Fraser, P. (2000) Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus. *Mol Cell* **5:** 377–386.

Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R., Fraser, P. (2000) Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus. *Mol Cell,* **5:** 377–386.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223-227.

Jeon, J.T., Carlborg, O., Tornsten, A., Guiffra, E., Amaerger, V., Chardon, P., et al., (1999) A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nature Genet.* **21:** 157–158.

Kaczynski, J., Cook, T., Urrutia, R. (2003) Sp1- and Krüppel-like transcription factors. *Genome Biol.* **4(2):**206.

Kaliman, P., Canicio, J., Testar, X., Palacin, M., Zorzano, A. (1999) Insulin-like growth factor-II, phosphatidylinositol 3-kinase, nuclear factor-kappaB and inducible nitric-oxide synthase define a common myogenic signaling pathway. *J Biol Chem* **274:** 17437-17444.

Kel, A.E., GöBling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., Wingender, E. (2003) MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.,* **13:** 3576-3579.

Kel, A.E., Kondrakhin, Y.V., Kolpakov, Ph.A., Kel, O.V., Romashenko, A., Wingender, E., Milanesi, L., Kolchanov, N.A. (1995) Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.,* **3:** 197–205.

Knüppel, R., Dietze, P., Lehnberg, W., Frech, K., Wingender,E. (1994) TRANSFAC® retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J. *Comput. Biol.,* **1:** 191–198.

Kokta, T. A., Dodson, M. V., Gertler, A., Hill, R. A. (2004) Intercellular signaling between adipose tissue and muscle tissue. *Domest Anim Endocrinol* **27:** 303-331.

Larson, G., Dobney, K., Albarella, U., Fang, M., Smith, E.M. et al., (2005) Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science,* **307:** 1618-1621.

Liang ,J., Song,W., Tromp,G., Kolattukudy P,E., Fu M .(2008) Genome-wide survey and expression profiling of CCCH-zinc finger family reveals a functional module in macrophage activation. *PLoS One* **3:** e2880.

Lin, R., Maeda, S., Liu, C., Karin, M., Edgington, T.S. (2007) A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene,* **26 (6):** 851–8.

Lin, R., Maeda, S., Liu, C., Karin, M., Edgington, T.S. (2007) A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene,* **26 (6):** 851–8.

Liu, J. P., Baker, J., Perkins, A. S., Robertson, E. J., Efstratiadis, A. (1993) Mice carrying null mutations of the genes encoding insulin-like growth factor I (Igf-1) and type 1 IGF receptor (Igf1r). *Cell* **75:** 59-72.

Macian F (2005) NFAT proteins: key regulators of T-cell development and function. *Nat. Rev. Immunol.* **5** (6): 472–84

Mancini-Dinardo, D., Steele , S.J., Levorse , J.M., Ingram, R.S., Tilghman , S.M. (2006) Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev,* **20:** 1268–1282.

Mancini-Dinardo, D., Steele, S.J, Levorse, J.M, Ingram, R.S., Tilghman, S.M. (2006) Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev* **20:** 1268–1282.

Marecki, S., Riendeau, C.J., Liang, M.D., Fenton, M.J. (2001) PU.1 and multiple IFN regulatory factor proteins synergize to mediate transcriptional activation of the human IL-1 beta gene. *J. Immunol.* **166:** 6829–6838.

Markljung, E., Jiang, L., Jaffe, J.D., Mikkelsen, T.S., Wallerman, O. et al., (2009) ZBED6, a Novel transcription Fcator Derived from Domesticated DNA Transposon Regulates IGF2 expression and Muscle Growth. *PLoS Biol,* **7(12):** e1000256.

Mattick, J.S, Amaral, P.P, Dinger, M.E, Mercer, T.R, Mehler, M.F. (2009) RNA regulation of epigenetic processes. *Bioessays,* **31:**51-59.

Mattick, J.S., Amaral, P.P., Dinger, M.E., Mercer, T.R., Mehler, M.F. (2009) RNA regulation of epigenetic processes. *Bioessays*, **31:**51-59

Matys, V., Fricke, E., Geffer, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V. et al. (2003) TRANSFAC: Transcriptional regulation, from patterns to profiles*. Nucleic Acids Res.* **31:** 374-378.

Mehler, M.F., Mattick, J.S. (2006) Non-coding RNAs in the nervous system. *J Physiol,* **575:**333-341.

Mehler, M.F., Mattick, J.S. (2006) Non-coding RNAs in the nervous system. *J Physiol*, **575:**333-341

Mercer, T.R., Dinger, M.E., Mattick, J.S. (2009) Long non-coding RNAs: insights into functions. *Nat Rev Genet,* **10:**155-159.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., Mattick, J.S. (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci USA*, **105:**716-721.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., Mattick, J.S. (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci USA*, **105:**716-721.

Milan, D., Jeon, J.T., Looft, C., Amerger, V., Robic, A., Thelander, M., Rogel-Gaillard, C. et al., (2000). A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* **288:**1248-1251.

Mitsuya, K., Meguro, M., Lee, M.P., *et al.* (1999) LIT1, an imprinted antisense RNA in the human KvLQT1 locus identified by screening for differentially expressed transcripts using monochromosomal hybrids. *Human Molecular Genetics,* **8 (7):** 1209–17.

Mitsuya, K., Meguro, M., Lee, M.P., *et al.* (1999) LIT1, an imprinted antisense RNA in the human KvLQT1 locus identified by screening for differentially expressed transcripts using monochromosomal hybrids. *Human Molecular Genetics,* **8 (7):** 1209–17.

Nesterova, T.B., Barton,S.C., Surani, M.A., Brockdorff, N. (2001) Loss of Xist imprinting in diploid parthenogenetic preimplantation embryos. *Developmental Biology* **235** (**2**)**:** 343–50.

Nesterova,T.B., Barton, S.C., Surani , M.A., Brockdorff, N. (2001). Loss of Xist imprinting in diploid parthenogenetic preimplantation embryos. *Developmental Biology,* **235 (2):** 343–50.

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. *Nature Genet.* **21:** 155–156.

Numoto, M., Niwa, O., Kaplan, J., Wong, K. K., Merrell, K., Kamiya, K., Yanagihara, K., Calame, K. (1993) Transcriptional repressor ZF5 identifies a new conserved domain in zinc finger proteins. *Nucleic Acids Res* **21 (16):** 3767-75.

Obata, T., Yanagidani, A., Yokoro, K., Numoto, M., Yamamoto, S. (1999) Analysis of the consensus binding sequence and the DNA-binding domain of ZF5. *Biochem Biophys Res Commun.***255(2):**528-34.

Oksbjerg, N., Gondret, F., Vestergaard, M. (2004) Basic principles of muscle development and growth in meat-producing mammals as affected by the insulin-like growth factor (IGF) system. *Domest Anim Endocrinol*, **27:** 219-240.

Oksbjerg, N., Gondret, F., Vestergaard, M. (2004) Basic principles of muscle development and growth in meat-producing mammals as affected by the insulin-like growth factor (IGF) system. *Domest Anim Endocrinol* **27:** 219-240.

Ong ,S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., et al. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1:** 376–386.

Orso, F., Penna, E. , Cimino, D., Astanina, E., Maione, F., Valdembri,D., Giraudo, E., Serini, G., Sismondi, P., De Bortoli, M., Taverna, D. (2008) AP-2 and AP-2 regulate tumor progression via specific genetic programs. *The FASEB Journal*, **22:** 2702-2714.

Owens, P. C., Gatford, K. L., Walton, P. E., Morley, W. ,Campbell, R. G. (1999) The relationship between endogenous insulin-like growth factors and growth in pigs. *J Anim Sci* **77**: 2098-2103.

Pang, K.C, Frith, M.C, Mattick, J.S. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics,* **22 (1):** 1–5.

Pang, K.C., Dinger, M.E., Mercer, T.R., Malquori, L., Grimmond, S.M., Chen, W., Mattick, J.S. (2009) Genome-wide identification of long noncoding RNAs in CD8+ T cells. *J Immunol,* **182:**7738-7748.

Pang, K.C., Dinger, M.E., Mercer, T.R., Malquori, L., Grimmond, S,M., Chen, W., Mattick , J.S. (2009) Genome-wide identification of long noncoding RNAs in CD8+ T cells. *J Immunol*, **182**:7738-7748.

Pang, K.C., Frith, M.C., Mattick, J.S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics,* **22 (1):** 1–5.

Penny G. D, Kay G. F, Sheardown S. A, Rastan S, Brockdorff ,N. (1996) Requirement for Xist in X chromosome inactivation. *Nature* **379:** 131–137.

Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., Brockdorff, N. (1996) Requirement for Xist in X chromosome inactivation. *Nature,* **379:** 131–137.

Pibouin, L., Villaudy, J., Ferbus, D., *et al.* (2002) Cloning of the mRNA of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. *Cancer Genetics and Cytogenetics,* **133 (1):** 55–60.

Pibouin, L., Villaudy, J., Ferbus, D., *et al.* (2002) Cloning of the mRNA of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas. *Cancer Genetics and Cytogenetics,* **133 (1):** 55–60.

Pickert, L., Reuter, I., Klawonn, F. and Wingender, E. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics* **14:** 244-251.

Ponjavic, J., Ponting, C. P., Lunter G (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17:** 556-565.

Prasanth, K. V., Spector, D. L. (2007) Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev* **21:** 11-42.

Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23:** 4878-4884.

Rinn , J.L., Kertesz, M., Wang, J.K., *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell,* **129 (7):** 1311–23.

Rinn, J.L, Kertesz, M, Wang, J.K., *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell,* **129 (7):** 1311–23.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004) JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32:** D91–D94.

Sleutels, F., Zwart, R., Barlow, D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* **415:** 810–813.

Sleutels, F., Zwart, R., Barlow, D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature,* **415:** 810–813.

Struhl, K (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology,* **14 (2):** 103–5.

Struhl, K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology* **14 (2):** 103–5.

Tao, Y., Kassatly, R., Cress, W.D., Horowitz, J.M. (1997) Subunit composition determines E2F DNA-binding site specificity. *Mol. Cell. Biol.,* **17:** 6994–7007.

Valenta, T., Lukas, J., Doubravska, L., Fafilek, B., Korinek, V. (2006) HIC1 attenuates Wnt signaling by recruitment of TCF-4 and beta-catenin to the nuclear bodies. *EMBO J.* **25(11):** 2326-37.

Van Laere, A.S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., et al. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature,* **425:** 832–836.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys,V., Michael, H., Ohnhaeuser, R. et al. (2001) The TRANSFAC system on gene expression regulation. *NucleicAcids Res.,* **29:** 281–283.

Wingender, E., Dietze, P., Karas, H., Knüppel, R. (1996) TRANSFAC®: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.,* **24:**238-241.

Wutz , A., Gribnau, J. (2007)X inactivation Xplained. *Current Opinion in Genetics & Development,* **17 (5):** 387–93.