Cis-acting Control of Human Developmental Regulator *GLI2*

By

RASHID MINHAS

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University

Islamabad, Pakistan

2015

Cis-acting Control of Human Developmental Regulator *GLI2*

By

RASHID MINHAS

A thesis

In the partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

In

BIOINFORMATICS

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University

Islamabad, Pakistan

2015

CERTIFICATE

This thesis submitted by **Rashid Minhas** is accepted in its present form by the National Center for Bioinformatics, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad as satisfying the thesis requirements for the degree of Doctor of Philosophy (Ph.D) in Bioinformatics.

Internal Examiner: _____

Dr. Amir Ali Abbasi

External Examiner: _____

Chairman: _____

Professor Dr. Wasim Ahmed

Dated: _____

To my parents and family

For their endless love, support and encouragement

ACKNOWLEDGEMENTS

I bow my head before Almighty Allah, the Lord of the worlds, The Omnipotent, The Gracious and The Merciful, who blessed me with good health and conducive environment and enabled me to work and complete my PhD work. Without Allah's divine help, I would not have been able to achieve anything in my life.

My special praise to Holy Prophet (SAW), the most perfect among all human beings ever born on the surface of the earth, from the deepest core of my heart who is forever a model of guidance and knowledge for the whole mankind and a source of knowledge and blessings for the entire creation. He has guided his Ummah to seek knowledge from cradle to grave and this has awakened in me the strong desire to undertake this work and write up.

I owe my sincerest gratitude to my supervisor, Dr.Amir, who has supported me throughout my thesis with his patience, valuable suggestions and expertise knowledge. I attribute the level of my PhD degree to his encouragement, keen interest, skillful guidance, effort and without his thesis, too, would not have been completed or written. One simply could not wish for a better devoted, ideal supervisor and scientist. I also take this opportunity to express my deepest gratitude and sincere thanks to Prof. Dr. Wasim Ahmed, Dean, Faculty of Biological sciences.

I am also pleased to Higher Education Commission of Pakistan for supporting and providing funds during my PhD work and IRSIP program. I wish to express sincere thanks to Dr.Greg Elgar (NIMR, London, UK) who has given me an opportunity to visit one of the best labs in the world and to work with his team members. I am grateful to Dr.Stefan Pauls, Dr.Laura Doglio, Dr.Boris, Joseph Grece and everyone in System biology lab, NIMR.

I express gratitude my parents for their love and support throughout my life. They sacrificed their own wishes to fulfill my dreams. No words in any dictionary of world for my parents, sisters and brothers Tahir and Junaid for their amazing love and support not only during my research work and studies but also in every step of my life. A special mention in this regard are my brothers in law Khurram and Khalil.

I convey my heartiest and sincerest acknowledgements to my wife Yusra Hasan Siddiqui for her love, care, moral support and kind cooperation. I have to mention here, how hard she worked to pull me into the scientific community. She wanted to see me in the scientific world, and so I worked hard to fulfill her wish. This thesis is only the beginning of my scientific journey.

In my daily work, I have been blessed with a friendly and cheerful group of fellow students: Nashaiman, Shahid, Rabail, Nazia and others, during my stay in EVOGENO lab. I wish to express sincere thanks to them for their utmost cooperation and their valuable company.

I am thankful to NCB staff: Mr.Talib, Mr.Ali, Mr.M.Naseer, Mr.Yasir and Mr.Naseer Ahmed for their kind cooperation.

I am also very pleased to thank my mentor Gulfaraz Khan for all the guidance and help since my childhood.

It would be very unjust if I do not mention my best friends: Riaz, Manazar Bhai, Sajid, Zeeshan, Abid, Hameed, Mudasar, and Saad for lending help in times of need.
In the end I want to present my unbending thanks to all those hands who prayed for my betterment and serenity.

R- Minhas

Contents

List of Figures i

List of Tables iii

List of Abbreviations iv

Summary vi

INTRODUCTION **1**

1.1 *Cis*-regulatory control of eukaryotic genes **1**

1.1.1 Promoter **3**

1.1.2 Distal regulatory elements **4**

1.2 Contribution of *cis*-acting regulatory elements in vertebrate development **8**

1.2.1 Role of *cis*-acting elements in tetrapod limb development **9**

1.3 Contribution of *cis*-acting elements in human diseases **10**

1.4 Contribution of *cis*-acting elements in animal evolution **13**

1.5 How to identify *cis*-acting regulatory sequences? **15**

1.5.1 Experimental techniques **15**

1.5.2 *In silico* identification: Comparative genomics to identify CNEs **16**

1.6 The GLI family: key developmental regulators **18**

1.7 Molecular roles of GLI family are highly complex and context dependent **21**

1.7.1 Gli2 is critical in Hh signaling for development of neural tube, limb and internal organs **24**

1.8 *Cis*-regulatory underpinning of GLI family **27**

1.9 Aims and objectives **28**

MATERIALS AND METHODS **30**

2.1 *In silico* protocols **30**

2.1.1 Sequence collection and comparative analysis **30**

2.1.2 *In silico* mapping of conserved TFBSs within each CNE **30**

2.1.3 Syntenic analysis **31**

2.2 DNA protocols **32**

2.2.1 Genomic DNA extraction from whole blood **32**

2.2.2 Ethanol precipitation to purify the DNA **33**

2.3 PCR Protocols **33**

2.3.1 Primer designing and dilution **33**

2.3.2 Polymerase chain reaction (PCR) **34**

2.3.3 DNA purification with gel extraction kit **35**

2.4 Bacterial Protocols 36

2.4.1 Preparation of media and agar plates 36

2.4.2 Preparing competent cells and storage 36

2.4.3 Transformations 37

2.4.4 Plasmid preps 38

2.5 Cloning CNEs into the transposon vector 38

2.5.1 Generation of entry clone 38

2.5.2 Restriction digestion of TOPO entry clone with *EcoRI* 39

2.5.3 Orientation screening of insert-TOPO vector clones 39

2.5.4 Generation of destination clone or recombination with pGW_*cfos*EGFP destination vector 39

2.6 *In vitro* transcription of transposase RNA 40

2.6.1 Capped transcription reaction assembly 41

2.6.2 Recovery of RNA by lithium chloride precipitation 41

2.7 Generation of transgenic zebrafish 42

2.7.1 Zebrafish breeding and mating 42

2.7.2 Needle pulling and ramp test 43

2.7.3 Co-injection reporter assay 43

2.7.4 *Tol2* mediated transgenesis 44

2.7.5 Post injection treatments and embryo screening 44

RESULTS 46

3.1 Identification of tetrapod-teleost conserved non-coding elements in *GLI2* locus by comparative sequence analysis 46

3.2 Identification of co-orthologs in the zebrafish genome 48

3.3 Association of identified human CNEs to *GLI2* locus by syntenic mapping 49

3.4 *In vivo* functional analysis of *GLI2*-associated CNEs using a co-injection assay in transiently transfected zebrafish embryos 50

3.4.1 CNE1 52

3.4.2 CNE2 53

3.4.3 CNE3 54

3.4.4 CNE4 55

3.4.5 CNE5 56

3.4.6 Zebrafish duplicated CNEs (CNE2a and CNE2b) 57

3.4.7 Comprehensive outline of GFP expression territories in zebrafish embryos at 24 hpf or 48 hpf 58

3.5 Study of *GLI2*-associated CNEs by *Tol2* based transgenic assay 59

3.5.1 CNE1 evokes GFP expression in epidermis with the *Tol2* reporter system .. 61

3.5.2 CNE2 induces reporter gene expression in CNS and pectoral fin 62

3.5.3 CNE3 governs GFP expression in CNS by *Tol2* 65

3.5.4 CNE4 induces GFP expression limited to muscle cells 66

3.5.5 CNE5 induced GFP expression in CNS and pectoral fin by *Tol2* transgenesis 67

3.5.6 Duplicated CNEs (CNE2a and CNE2b) have similar GFP expression in hindbrain and pectoral fin with *Tol2* transgenesis 68

3.6 *In silico* mapping of conserved transcription factor binding sites (TFBSs) within each CNE 69

DISCUSSION 74

4.1 Regulation of Hh mediator is crucial for vertebrate embryogenesis 75

4.2 Evolutionary sequence comparison reveals candidate *GLI2* enhancers 76

4.3 Progressive expansion of novel regulatory components around an ancient enhancer element 79

4.4 *GLI2*-associated CNEs show tissue-specific regulatory activity *in vivo* 80

4.5 CNE2, CNE3, and CNE5 induce reporter expression that coincides with known sites of GLI2 activity in CNS 84

4.6 *Cis*-regulatory control of GLI2 expression in developing pectoral fin 87

4.7 *Cis*-regulatory control of *GLI2* expression in skin cells 88

4.8 CNE1, CNE3 and CNE5 activate reporter expression within cardiac chamber and circulatory blood cells 89

4.9 CNE3 activity in branchial arch and otic vesicle 89

4.10 *GLI2*-associated CNEs appear to drive overlapping GFP expression in zebrafish muscle 90

4.11 Duplicated CNEs suggest overlapping expression pattern in hindbrain and pectoral fin 91

4.12 Conclusions and future perspectives 92

REFERENCES 94

APPENDIX 105

PUBLICATIONS 110

List of Figures

Figure 1. 1 Schematic of a typical gene regulatory region 2
Figure 1. 2 An overview of gene regulation by distant acting enhancers 5
Figure 1. 3 Evolution in action: a genetic basis. 14
Figure 1. 4 Comparison of mammals-fugu genome revealed highly conserved non-coding sequences 17
Figure 1. 5 Hh-signaling cascade is highly conserved among insects and vertebrates 19
Figure 1. 6 Schematic illustration of domains and motifs in Gli proteins. 21
Figure 1. 7 RNA in situ hybridization studies showing gli2a and gli2b expression in several embryonic domains at 24 and 48hpf *(Source: www.zfin.org)* 26

Figure 3. 1 MLAGAN alignment of the genomic region encompassing *GLI2* 47
Figure 3. 2 Comparative genic architecture of human *GLI2*-associated CNEs in zebrafish revealed CNE2 is duplicated in zebrafish genome. 49
Figure 3. 3 Human *GLI2*-associated CNEs maintain their physical linkage with GLI2 gene .. 50
Figure 3. 4 Schematic representation of co-injection strategy used for *in vivo* characterization of CNEs 52
Figure 3. 5 CNE1 mediated reporter gene expression prediminantly in brain and ventral caudal region 53
Figure 3. 6 CNE2 exclusively regulates GFP expression in notochord at day-2 and at day-3 54
Figure 3. 7 CNE3 mediated reporter gene is expressed predominantly in otic vesicle, hindbrain and muscle cells. 55
Figure 3. 8 CNE4 specifically regulates GFP expression in muscle cells at day-2 and at day-3 56
Figure 3. 9 CNE5 drives GFP expression in CNS. 57
Figure 3. 10 Sites of GFP expression induced by duplicated orthologous sequences of CNE2 in zebrafish embryos. 58
Figure 3. 11 Sites of GFP expression induced by *GLI2*-associated CNEs in zebrafish embryos 59
Figure 3. 12 Transient expression assay in zebrafish. 60
Figure 3. 13 Comparison of GFP expression in epidermis by co-injection and *Tol2* transgenic system. 61
Figure 3. 14 CNE2 drives GFP expression mainly in the spinal cord, epidermis and notochord by *Tol2* transgenic system. 62
Figure 3. 15 CNE2 drives GFP expression in the hindbrain and fin by *Tol2* reporter system 64
Figure 3. 16 Truncated *GLI2* CNE2 (208bp) triggers GFP expression in notochord and hindbrain. 64
Figure 3. 17 CNE3 controls GFP expression in the central nervous system and sensory organs of zebrafish embryos. 66
Figure 3. 18 CNE4 induces non-significant GFP expression by *Tol2* transgenic system. 66

Figure 3. 19 CNE5 mediated reporter gene expression was more prominent in hindbrain and spinal cord. 68
Figure 3. 20 Dr-*gli2*_CNE2a and Dr-*gli2*_CNE2b mediated reporter gene expression remained more prominent in hindbrain and pectoral fin 69
Figure 3. 21 Multiple alignments show a human/fish core conserved sequence track within CNE1 70
Figure 3. 22 Graphical representation of transcription factors binding motifs identified by MEME 71
Figure 3. 23 Conserved transcription factor binding sites in CNE1. 72

Figure A1. Map shows the features of the pCR™8/GW/TOPO® vector. 105
Figure A2. Map of the *cfos-IsceI*-EGFP plasmid. 106
Figure A3. Map of the *β-globin* EGFP plasmid. 107
Figure A4. Map of the pCS-TP vector. 108

List of Tables

Table 1.1 List of some human *cis*-ruption diseases..... 12

Table 2. 1 Primers used to amplify the intra-*GLI2* conserved non-coding elements (CNEs) for co-injection assay and cloning into *Tol2* vector 35

Table 3. 1 Tetrapod-teleost conserved non-coding elements (CNEs) from Intron of human *GLI2* selected for functional analysis in transgenic zebrafish assay 48
Table 3. 2 Computationally predicted transcription factors binding sites and their associated transcription factors 73

Table 4. 1 Summary of studies identifying GLI morphopathies 76
Table 4. 2 Comparison of reporter gene expression induced by intra-*GLI2* CNEs 83
Table 4. 3 Reported endogenous expression pattern of Gli2 and gli2a/gli2b in vertebrates 84

Table A1 1 Plasmid-specific primers and their annealing temperatures..... 109

List of Abbreviations

°C	Degree Celsius
µg	Micro gram
µl	Microlitre
AP	Anterior posterior
bp	base pair
BLAST	Basic local alignment search tool
Cl	Chloride
CNE	Conserved non coding elements
CNS	Central nervous system
CRM	*cis*-regulatory module
dCNE	Duplicated conserved non coding element
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
Dr	Danio *rerio*
DV	Dorso ventral
E.*coli*	Escherichia coli
EDTA	Ethylene diamine tetra acetic acid
EE	Expressing embryos
EM	Embryo medium
GFP	Green fluorescent protein
HCl	Hydrochloric acid
Hh	Hedgehog
hpf	Hour post fertilization
Kb	kilo base pair
LAGAN	Limited area global alignment of nucleotide

MEME	Multiple Expectation Maximization for Motif
Mg	Magnesium
mg	Milli gram
ml	Millilitre
*m*RNA	*Messenger* ribonucleic acid
MSA	Multiple sequence alignment
Mya	Million years ago
NaCl	Sodium chloride
NCBI	National centre for Biotechnology and information
ng	Nanogram
PBS	Phosphate buffer saline
PCR	Polymerase chain reaction
PD	Proximo distal
PTU	Phenylthiourea
q	Long arm of chromosome
RFP	Red fluorescent protein
RNA	Ribonucleic acid
rpm	Revolution per minute
Shh	Sonic hedgehog
Taq	*Thermus aquaticus*
TBE	Tris borate EDTA
TBE	Tris borate ethylene
TE	Tris EDTA
TF	Transcription factor
TFBS	Transcription factor binding site
UCSC	University of California, Santa Cruz

Summary

Background

Gene regulation is a highly complex process and involves the coordination of RNA polymerase, enhancers and promoters, multiple transcription factors, silencers, insulators, and locus control regions. The spatiotemporal activity of a gene requires the presence of intact coding sequences as well as properly functioning *cis*-acting regulatory control. The identification and functional characterization of *cis*-acting elements is one of the greatest challenges of the post-genomic era for better understanding of the language, syntax and grammar that is encoded in regulatory DNA. Gene expression plays a significant role in the evolution of vertebrate complexity and diversity during development. The vertebrate GLI family (glioma-associated oncogene family members 1, 2, and 3) of transcription factors are key transducers of one such pathway known as Sonic hedgehog (Shh) signaling. Shh-Gli interactions have been extensively scrutinized from last couple of decades by genetic, molecular and biochemical means. In the present study, an important mediator of Shh signaling, *GLI2* gene was focused. GLI2 has been implicated in diverse set of embryonic developmental processes including patterning and growth of the central nervous system, craniofacial structures, skeleton, limb, skin, and internal organs. In addition, a number of clinical conditions and developmental defects have already been associated with mutation in this key developmental regulator. Precise spatio-temporal expression of Gli2 is critical for the proper specification of these structures in vertebrates. In this study, I identified and characterized, the *cis*-regulatory catalogue of human GLI2 gene.

Results & Conclusions

Towards the elucidation of genetic mechanisms, by which the transcription of *GLI2* genes is regulated during early embryonic development, deep evolutionary constraints (tetrapod-teleost) have been used as an indicator to pinpoint the conserved intronic intervals of human *GLI2* gene. This ancient catalogue act as tissue-specific enhancers in transient transgenic zebrafish assays and induce reporter gene expression in a diverse set of embryonic domains, where GLI2 is known to be expressed endogenously. Interestingly, these *GLI2*-associated enhancers have considerable overlapping expression territories during zebrafish development.

Significance

Elucidation of the *GLI2*-associated *cis*-regulatory network offers a novel perspective for better understanding the genetic mechanisms by which the downstream effectors of Hh signaling cascade might be regulated during embryogenesis. In addition, these *cis*-regulatory modules are strong candidates for mutational studies of *GLI2*-associated human birth defects, which cannot be attributed to any exonic mutation.

INTRODUCTION

Transcriptional regulation of gene expression is said to play an important role in establishing morphological diversity in phenotype and biological functions from a common set of genes. With an increasing body of emerging evidences, it is now widely accepted that gene regulatory networks drive increased complexity in animal evolution (Levine, 2010, Levine & Tjian, 2003). However, gene regulation mechanisms differ markedly in prokaryotes and eukaryotes. In prokaryotes, gene regulation is simplistic as co-regulated genes are often ordered into operons on adjacent loci, to be transcribed collectively through a single promoter region (Jean-Jack, 2010). For example, *E.coli lac* operon is a cluster of three genes involved in lactose metabolism. It is found that this operon is dual regulated via a single region, such that it is inhibited by the lac repressor within the promoter, and activated by lactose binding to repressor (Jean-Jack, 2010). It is due to this co-regulation of gene clusters, that prokaryotes have simpler physiology as compared to eukaryotes. Contrary to the genomic organization of many prokaryotic organisms that are compact and gene rich, eukaryotes have evolved to involve a more complex and combinatorial set of regulation of transcription.

1.1 *Cis*-regulatory control of eukaryotic genes

Transcriptional complexity in regulation significantly increases from prokaryotic to simple, single-cell, eukaryotic organisms, and then again increases in complex metazoan eukaryotes. Metazoan genes hold extreme intricate regulatory sequences that direct complex organismal body organization. Precise spatial and temporal

expression of a gene in metazoans is controlled by *cis*-acting elements. These non-coding DNAs act as *cis*-regulators, and are often scattered over great distances or within an intron of the gene they control (Lettice *et al.*, 2003, Levine & Tjian, 2003, Venkatesh *et al.*, 1996). Eukaryotic c*is*-regulatory elements are typically divided into two major classes: promoters and distal regulatory elements (Figure 1.1). Distal regulatory elements being composed of enhancers, silencers, insulators, and locus control regions. These *cis*-acting elements contain sites for *trans*-acting DNA-binding transcription factors, which function either to enhance or repress transcription.

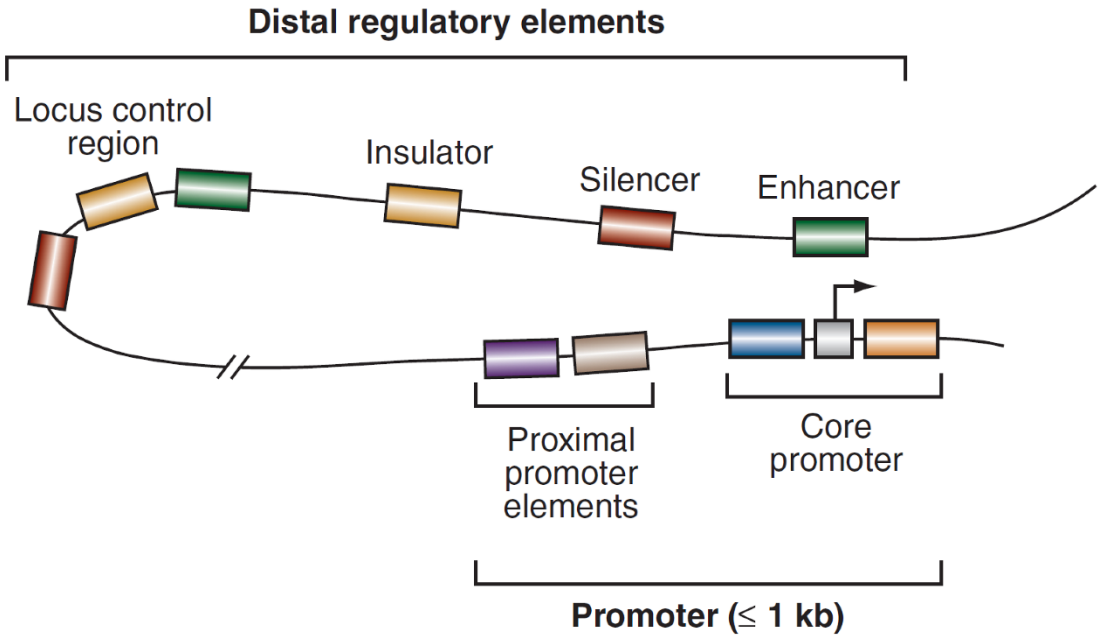

Figure 1. 1 Schematic of a typical gene regulatory region

The promoter, which is composed of a core promoter and proximal promoter elements typically spans less than 1 kb pairs. Distal (upstream) regulatory elements, which can include enhancers, silencers, insulators, and locus control regions, can be located up to one Mb from the promoter. These distal elements may contact the core promoter or proximal promoter through a mechanism that involves looping out the intervening DNA. (Adapted and modified from Matson et al. 2006).

1.1.1 Promoter

Promoter region can be generally defined as a minimal stretch of contiguous DNA sequences required for the initiation of transcription (Maston *et al.*, 2006). Most of the eukaryotic genes contain a single promoter located near the transcription start site. However, some genes contain alternative promoters that activate transcription at special positions in the genome (Cooper *et al.*, 2006). The gene promoter region is generally divided into two sub-classes, i.e. basal (core promoter) and proximal promoter.

1.1.1.1 Core promoter

Usually, the core promoter includes the site of transcription initiation and there are several sequence motifs that are commonly found in core promoters which include the TATA box (binding site for TATA-binding protein (TBP)), Initiator, TFIIB recognition element (BRE), downstream core promoter element (DPE), and motif ten element (MTE) (Butler & Kadonaga, 2002, Birney *et al.*, 2007). A particular core promoter may contain some, all, or none of these elements. The core promoter acts as a docking site for the assembly of basic transcriptional machinery and pre initiation complex (PIC) that carries TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, and RNA polymerase II (Pol II), which all act mutually to specify the transcription start site (Thomas & Chiang, 2006).

1.1.1.2 Proximal promoter

Proximal promoter is positioned immediately upstream or downstream from the core promoter at the 5’ end, and is usually within from −250 to +250 nucleotide upstream of the start of transcription (Butler & Kadonaga, 2002). Proximal promoter

typically contains multiple binding sites for transcription factors, and increases the frequency of initiation of transcription, when positioned near the transcriptional start site (Levine & Tjian, 2003). The transcription factors that bind promoter proximal elements do not always directly activate or repress transcription. Instead, they might serve as "tethering elements" that recruit long-range regulatory elements, such as enhancers, to the core promoter.

1.1.2 Distal regulatory elements

1.1.2.1 Enhancers

Enhancers are short intervals of DNA that can vary in size from as short as 100bp to a few kilo base pairs in length. They act as operational platforms to recruit multiple TFs through a cluster of short (6-12 bp) transcription factor binding sites (TFBSs), to regulate transcription of genes (Levine, 2010, Paparidis *et al.*, 2007, Abbasi *et al.*, 2007, Pennacchio *et al.*, 2006). Enhancers activate gene expression independent of their orientation and are scattered across the non-protein coding genomic intervals, as reviewed by (Pennacchio *et al.*, 2013). Enhancers are present upstream or downstream of a gene, within an intron and/or UTR regions, in the intron of an unrelated gene, or maybe docked more than one mega base pairs away from their target gene promoters (Abbasi *et al.*, 2007, Lettice *et al.*, 2003, Levine, 2010, Pennacchio *et al.*, 2006, Woolfe *et al.*, 2005). Moreover, an enhancer may regulate the expression of its target gene located on a different chromosome (Lomvardas *et al.*, 2006). Several models proposed that enhancers and core promoters are brought into close proximity by looping out the intervening DNA (Vilar & Saiz, 2005). Transcriptional enhancers determine where, when and how much a protein-coding

gene is expressed. A typical feature of enhancers is the modular mode by which they regulate gene expression, which means the binding of tissue-specific transcription factors to their respective enhancer is important in causing tissue-specific expression (Figure 1.2).

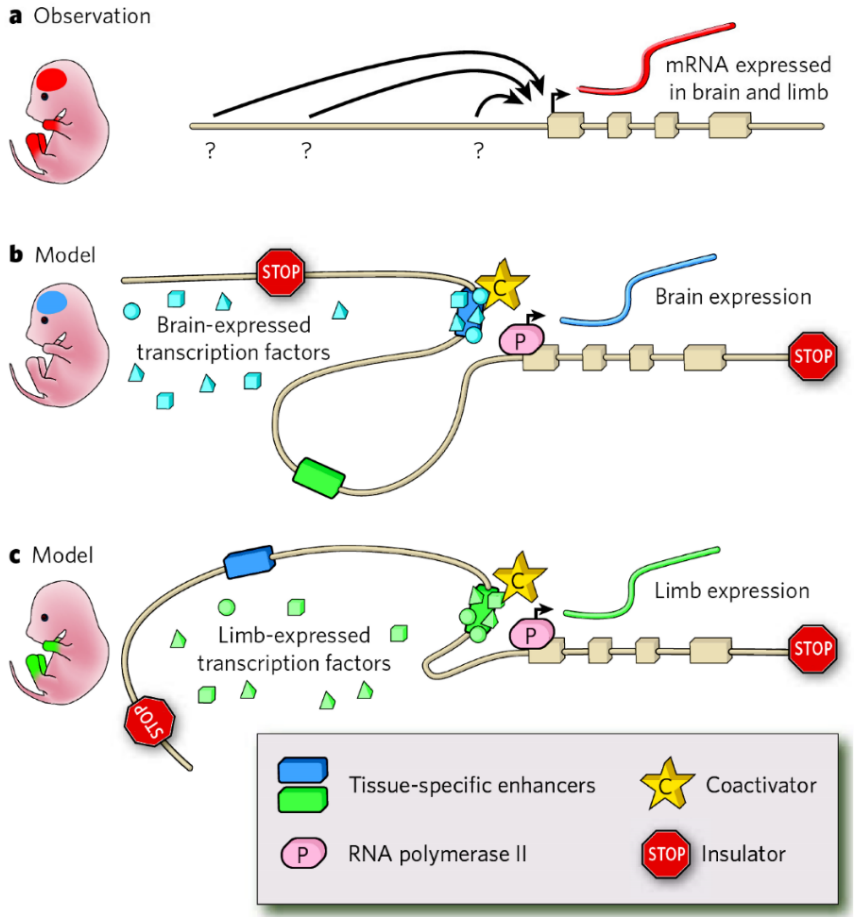

Figure 1. 2 An overview of gene regulation by distant acting enhancers

(a) An example of a gene (red) expressed in diverse developmental domains (brain and in limbs). Its activity depends on distant-acting and cis-regulatory modules. (b & c): Tissue-specific enhancers are thought to contain binding sites for multiple TFs. Only when all required TFs are present in a tissue does the enhancer become active: it binds to transcriptional co-activators, relocates into physical proximity with the gene promoter and activates transcription by RNA polymerase II (through a looping mechanism). Adapted and modified from (Visel et al., 2009).

1.1.2.1.1 Super enhancers

Recently, a subset of enhancers have been identified to be evolved in mammalian genomes, termed as 'super enhancers' which have crucial functions in defining cell identity (Whyte *et al.*, 2013). These large clusters of enhancers, up to 50 kb in size, are occupied by master transcription factors, and super-enhancer associated genes are generally expressed at higher levels than genes associated with typical enhancers. Super-enhancers are occupied by a large portion of the enhancer-associated RNA polymerase II and its associated cofactors and chromatin regulators, which can explain how they contribute to high-level transcription of associated genes. (Hnisz *et al.*, 2013) have generated a catalog of super-enhancers in 86 human cell and tissue types. It is found that disease-associated sequence variation is enriched in super-enhancers, and they have been investigated to play a role in cancer, where they are linked to critical oncogenic tumor genes (Loven *et al.*, 2013, Whyte *et al.*, 2013). Thus, super-enhancers provide biomarkers for disease diagnosis and therapy.

1.1.2.2 Silencers

Silencers are sequence-specific elements that suppress gene expression independent of their orientation and position from the promoter. Silencers can be present distant from their target gene, in its 3´-untranslated region, or in its intron (Levine & Tjian, 2003, Maston *et al.*, 2006). They may be short-range (acting at a distance of ~100bp) or long-range (acting at a distance of more than few kilo bases) silencers.

Although the precise mechanism of silencers is not fully elaborated several proposed theories of how silencers repress transcription is reviewed by (Maston *et al.*, 2006).

Silencers can act directly by binding adjacent transcriptional activators that interact with the promoter or by directly competing for the similar site. Silencers/repressors may stop access of enhancer-activator complexes to the promoter by repressing chromatin structure or may inhibit transcription by hindering PIC assembly (Chen & Widom, 2005).

1.1.2.3 Insulators

The term 'insulator' defines DNA elements which prevent the undesirable interaction of enhancers with promoter or a fence to condensation of the heterochromatin, thereby preventing genes from being affected by the transcriptional activity of neighboring genes. Insulators are ~300bp to 2kb in length and are believed to be an essential part of the regulatory machineries that ensure appropriate transactions between enhancers and the promoters (Levine & Tjian, 2003). The exact mechanism of an insulator's mode of action is not known. However, it is proposed that insulators have two main characteristics: (I) they can block communication between a promoter and an enhancer (enhancer-blocking insulators), and (II) can block the spread of repressive chromatin (barrier insulators) (Gaszner & Felsenfeld, 2006). Chicken β-globin insulator 5'HS4 is a well-known example of insulators in vertebrates (Felsenfeld *et al.*, 2004).

1.1.2.4 Locus control regions

Locus control regions (LCRs) are *cis*-acting DNA regions required for the activation of a gene cluster or an entire locus. LCRs were first identified in the human β-globin locus (Hardison *et al.*, 1997). They are composed of multiple *cis*-acting elements including insulators, silencers, enhancers, etc., which are bound by their specific

transcription factors and each of this differentially affects gene expression. Strong and specific enhancer activity is the characteristic feature of LCR. Several studies have identified LCR regions, which revealed that like enhancers and repressors, LCRs can regulate gene expression independent of their distance. LCRs are normally positioned upstream of their target gene. However, they can be present within an intron or downstream of a gene. LCRs can be highlighted by clusters of neighboring DNase I hypersensitive sites, and are supposed to be an open-chromatin domain for genes to which they are related. Mammalian β-globin LCR was among the first LCRs to be identified, and is also the best one studied, as reviewed by (Chakalova *et al.*, 2005). The human β-globin LCR lies approximately 6-25 kb upstream of the gene cluster locus which consists of five genes that are differentially expressed during development and are organized in order of their developmental expression. Mouse β-globin LCRs are orientation-dependent and inverting the LCR abolishes much of their role (Tanimoto *et al.*, 1999). LCRs can interact with their target genes with a looping mechanism similar to the mode of action of enhancers.

1.2 Contribution of *cis*-acting regulatory elements in vertebrate development

Development of a fertilized egg into an organism encompasses several critical events, and integral to this process is the precise and dynamic transcriptional regulation of developmental genes (Carroll 2005). Most enhancers are enriched around developmental regulators, and evolutionarily conserved sequences. The expression of developmental genes is regulated by multiple, modular enhancers, and alterations in the expression pattern of a gene may evolve through modifications in

cis-regulatory sequences, or in the deployment and activity of the transcription factors that control gene expression, or both.

Deletions or mutations in regulatory regions can result in developmental defects, such as deletion in brain-specific enhancer of *Otx2* in compound heterozygous embryos displays abnormal brain development (Kurokawa *et al.*, 2004). Similarly, deletion of the 200 bp enhancer of the *Hoxc8* gene results in delayed expression of the Hoxc8 protein, and various skeletal defects (Juan & Ruddle, 2003), and deletion of *Hand2* enhancer in mice results in defects in craniofacial development, including cleft palate and mandibular hypoplasia (Yanagisawa *et al.*, 2003). Interestingly, studies have shown that deletions or mutations in regulatory elements are restricted to defects in tissue-specific expression. Thus, *cis*-regulatory elements play a vital role in normal embryogenesis, by coordinating gene expression (Visel *et al.*, 2007).

1.2.1 Role of *cis*-acting elements in tetrapod limb development

A growing body of evidence from comprehensive analysis of limb key regulators like *HoxD* cluster, Sonic hedgehog (*Shh*), and *GLI3* propose that progress of regulatory components are the key for origin and subsequent morphological diversification of tetrapod appendicular skeleton (Spitz, 2001, Lettice *et al.*, 2003, Abbasi *et al.*, 2010). *HoxD* gene transcripts *(1-9)* are expressed in a time-dependent manner, early and throughout the limb bud, whereas *Hoxd10-13* are restricted towards the posterior of the early limb bud (digit-forming territory), and are essential for vertebrate limb buds. So, one of the critical genes controlling the regulatory landscape of vertebrate limb development is the 5'*HoxD* genes.

Temporal collinearity and progressive restriction of *HoxD* transcripts towards the posterior of early limb bud is controlled by two distinct poorly defined *cis*-regulatory modules: Early limb control region (ELCR) and Posterior Restriction (POST), positioned telomeric and centromeric respectively to the *HoxD* cluster, as reviewed by (Abbasi, 2011). ELCR induces transcription in a time-dependent manner, and POST imposes spatial restriction on *5'-HoxD* genes. ELCR-POST-induced nested pattern of *5'-HoxD* genes in posterior limb bud. This restricted localized expression induces posterior localized expression of *Shh* through interaction of HoxD 11-13 products with ZRS (zone of polarizing activity regulatory sequence), which subsequently triggers a second wave of *5'HoxD* genes to the limb bud. This *Shh* mediated anterior-to-posterior asymmetry in the *5'HoxD* genes, is translated into the anterior-posterior (AP) polarity of tetrapod limbs. Transgenic mice studies revealed that co-expression of *HoxD* in the digit forming territories during the second wave is governed by a region present at least 250 kb upstream to the *HoxD* cluster, famously known as GCR (Global control region). This region consists of several highly conserved intronic intervals (Spitz *et al.*, 2003).

Thus, it can be concluded that distinct *cis*-regulatory underpinnings (ELCR-POST and GCR) triggers coordinated spatio-temporal expression of a similar set of developmental regulators to form distal skeletal elements of the mature limb, from an initially homogeneous early limb bud (Abbasi, 2011).

1.3 Contribution of *cis*-acting elements in human diseases

Genetic diseases can be the result of coding mutation or deletion, or interference with normal gene expression through disruption of its *cis*-regulatory control. There

are many well documented pathological conditions having no coding mutation and may be due to disruption of communication between a gene and its associated *cis*-acting elements. Despite the argument of importance of non-coding DNA sequences by evolutionary biologists, the disease relevance of *cis*-acting mutations has largely been ignored in the past because of lack of insight into the identification of regulatory elements in non-coding regions. According to April 2009 statistics compiled by the Human Gene Mutation Database, 1459 regulatory mutations have been identified in over 700 genes that cause human-inherited disorders (Epstein, 2009).

Cis-regulatory mutations are less frequent than coding mutations, which generally disrupt the transcriptional process, and affect a broad range of morphological, physiological and neurological phenotypes. *Cis*-regulatory sequences harbor degenerate binding sites for multiple *trans*-regulatory sequences encoding transcription factors (Stern & Orgogozo, 2008). A change of single nucleotide within *cis*-regulatory modules can potentially modify the binding affinity for existing transcription factors, while removal or insertions can change the site spacing, delete existing binding sites or create novel ones (Wittkopp & Kalay, 2012). Mutations in *cis*-regulatory elements, or rearrangements affecting the position of distal regulatory elements with respect to their target genes, are associated with several pathological conditions (Table 1.1).

Eight possible mechanisms as explained by Kleinjan and Coutinho (2009), by which a *cis*-acting element can be disrupted from its target gene are: ‘’(1) deletion of long range *cis*-acting elements, (2) separation of *cis*-acting regulatory elements from the

gene promoters through chromosomal translocations or inversions, (3) deleterious mutations in *cis*-regulatory module, (4) disruption of the normal interactions between promoters and *cis*-acting enhancers through appearance of a new promoter, (5) alteration of local chromatin structure through interference by antisense transcripts, (6) disturbance of regional chromatin structure, (7) duplication of a *cis*-regulatory region, and (8) acquisition of inappropriate tissue-specific upstream promoters or enhancers'' (Kleinjan & Coutinho, 2009). In addition, many sequence variations within *cis*-acting regulatory modules potentially imposes acceptable effects on their activity resulting in incremental variations in spatio-temporal expression pattern of the associated gene, and can work as a fuel for evolution in phenotypic (especially morphological) divergence and complexity in several ways.

Table 1.1 List of some human *cis*-ruption diseases, adapted and modified from (Maston *et al.*, 2006)

Regulatory Element	**Disease**	**Mutation**	**Affected Gene**
Core Promoter	β-thalassemia	TATA box, CACCC box, DCE	β –globin
Proximal Promoter	Bernard-Soulier Syndrome	133bp upstream of TSS(GATA-1)	Gplb β
	Hemophilia	CCAAT box(C/EBP)	Factor *IX*
	Familial combined hyperlipidemia	39 bp upstream of TSS (Oct-1)	Lipoprotein lipase
	δ-thalassemia	77 bp upstream of TSS (GATA-1)	δ-globin
Enhancer	Preaxial polydactyly	1Mb upstream gene	SHH

	Van Buchem disease	Deletion~35kb downstream of	Sclerostin
	X-linked deafness	Microdeletions 900kb upstream	POU3F4
Silencer	Asthma and allergies	509 bp upstream of TSS (YY1)	TFG-β
	Fascioscapulohumeral muscular dystrophy	Deletion of D4Z4 repeats	4q35 gene
Insulator	Beckwith-Wiedemann syndrome	CTCF binding site (CTCF)	H19.Igf
LCR	α-thalassemia	62 kb deletion upstream of gene	α-globin genes

1.4 Contribution of *cis*-acting elements in animal evolution

Variations within *cis*-acting regulatory elements can work as a fuel for evolution in several ways. For instance, innovation of an enhancer around a developmental regulator can induce its expression in domains where it was not previously expressed (Shubin *et al.*, 2009). This expansion of *cis*-acting regulatory contents can potentially broaden the functional territories of the associated coding regions. Therefore, this expansion creates more complex developmental compartments and phenotypic evolution, through pleiotropy in the usage of the existing genetic toolkit. Sequence changes of individual transcription factor binding sites within enhancer elements can potentially alter the binding affinities for existing transcription factors. Similarly, modifications in the molecular anatomy of enhancers (enhancer structure) can create morphological diversity without affecting overall phenotype and fitness of an organism (Carroll, 2008, Bolker, 2000).

One such example of enhancer-mediated evolution is stated by (Shapiro *et al.*, 2004), wherein studies have found that variation in expression of the *Pitx1* gene underlies reduction of a large spine in the pelvic fin in three spine sticklebacks. The regulation of *Pitx1* activity has been partitioned into discrete regulatory elements, each of which controls this gene in a particular tissue; and changes in *Pitx1* expression result in pelvic-fin reduction (Figure 1.3). Shapiro *et al.* shows that marine sticklebacks with intact *Pitx1* regulatory elements develop robust pelvic spines, whereas mutations that specifically affect hind limb *Pitx1* expression alter the pelvic-fin structures of freshwater species, whereas other *Pitx1*-dependent structures are unaffected (Figure 1.3).

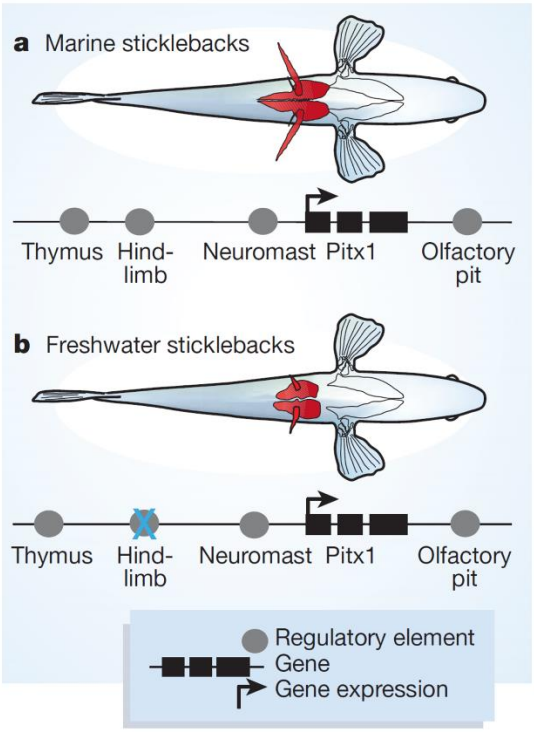

Figure 1. 3 Evolution in action: a genetic basis.
Variation in expression of the Pitx1 gene causes variation in decrease of a large spine in the pelvic fin in three spine sticklebacks. The regulation of Pitx1 activity has been partitioned into discrete regulatory elements, each of which controls this gene in a particular tissue; changes in Pitx1 expression that result in pelvic-fin reduction can thus be uncoupled from the requirements of other parts of the body for Pitx1 activity. Adapted and modified from (Shubin & Dahn, 2004)

1.5 How to identify *cis*-acting regulatory sequences?

1.5.1 Experimental techniques

As regulatory information is encoded in DNA sequences, and there is no known syntax for *cis*-regulatory regions, predicting and identifying these elements in the genome is a difficult task. Classical approaches were mostly restricted to random cloning of minimal promoter or promoter-proximal region, and subsequent deletion mapping to refine the critical region through a cell-based reporter assay, as reviewed by (Woolfe & Elgar, 2008). However, recently it has been found that functionally active genomic intervals are more susceptible to DNase degradation, and *cis*acting regulatory elements are observed to be associated with DNaseI hypersensitive sites. Thus, identification of degraded and retained genomic regions by subsequent Southern blot, PCR, microarray hybridization or sequencing, helps in characterizing these transcriptional regulatory elements (Pennacchio & Rubin, 2001). Another approach to identify *cis*-acting elements *in vivo,* has also been identified through BAC targeted enhancer trap approaches in transgenic zebrafish and mice (Durick *et al.*, 1999, Ellingsen *et al.*, 2005). Chemical modification, cross-linking studies like chromatin immunoprecipitation (ChIP)-ChIP and other gel-shift assays allow the determination of the entire spectrum of *in vivo* binding sites for a given protein. This helps in determining the sequence of transcription factor binding sites. Most of the *in vitro*/*in vivo* approaches are unguided, and thus are extremely laborious and time consuming (Pillai & Chellappan, 2009).

1.5.2 *In silico* identification: Comparative genomics to identify CNEs

Cis-regulatory sequences tend to be more conserved among different species than non-functional non-coding sequences (Elgar & Vavouri, 2008). Comparing DNA sequences of two or more distantly species provides a means of identifying conserved signatures that may have functional significance (Abbasi *et al.*, 2007, Pauls *et al.*, 2012). Aligning and comparing DNA sequences of distantly related vertebrate species is widely used to identify conserved non-coding elements (CNEs) in human and other genomes (McEwen *et al.*, 2006, Woolfe *et al.*, 2005, Woolfe & Elgar, 2007, Woolfe *et al.*, 2007). There are hundreds of thousands of non-coding sequences that are highly conserved between distantly related species and frequently clustered around developmental regulators (Figure 1.4). For example, human and Japanese pufferfish (*Fugu rubripes)* pair-wise genome comparison revealed several non-coding intervals that are nearly identical in sequence between human and fish that last shared a common ancestor more than 450 million years ago (Woolfe *et al.*, 2005, Woolfe *et al.*, 2007, Abbasi *et al.*, 2013, Abbasi *et al.*, 2010).

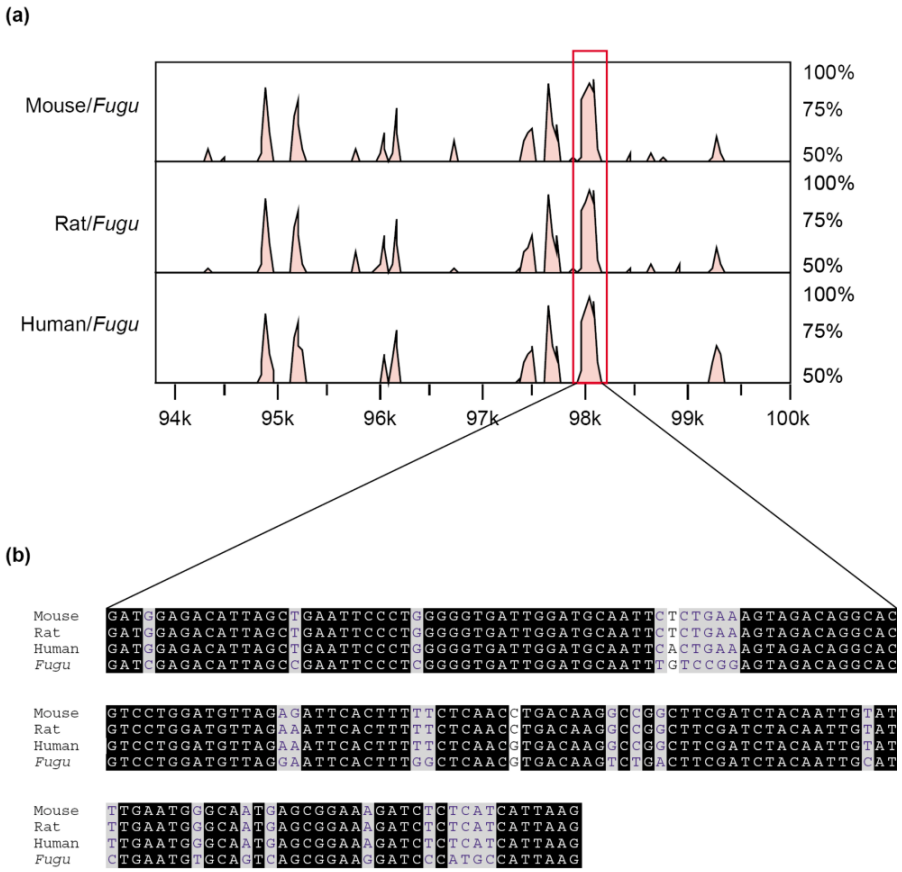

Figure 1. 4 Comparison of mammals-fugu genome revealed highly conserved non-coding sequences

(a) Vista tool displaying the conservation peaks among distantly related vertebrate species, (b) and DNA sequence similarity.

Comparisons of the human genome with phylogenetically close genomes like rodents or distantly related species like fish, have several advantages and disadvantages (Prabhakar *et al.*, 2006). For example, the human and mouse pairwise comparison detects huge degree of non-coding sequence similarities, probably due to an uneven rate of evolution across the genomes and short divergence time between human and mouse (last sharing a common ancestor ~75Mya) (Elgar & Vavouri, 2008). This background conservation makes it hard to distinguish neutrally evolving non-coding elements from functionally constrained ones (Pennacchio *et al.*, 2006, Visel *et al.*, 2008). Similarly, sequence comparison at extreme phylogenetic

distance (human/fish) has limitations to expose only the very specific sets of regulatory components that are significant for vertebrate embryogenesis and physiology. In contrast, such regulatory modules may be skipped in human/fish comparison that evolved after the divergence of fish and are required only for specific anatomical details to tetrapod and mammalian lineage (Abbasi *et al.*, 2010).

Several online libraries, for example JASPAR (http://jaspar.genereg.net/) and TRANSFAC (http://www.gene-regulation.com/pub/databases.html) are compiled to store the binding profiles of well-characterized and experimentally verified TFs. The availability of these libraries provides an unprecedented opportunity to decipher the non-coding DNA of vertebrate genomes for their *cis*-acting regulatory activity.

1.6 The GLI family: key developmental regulators

Hedgehog (Osorio *et al.*) signaling is shown to be highly conserved among insects and vertebrates, as reviewed by (Hui & Angers, 2011) (Figure 1.5). Shh is a secreted protein that undergoes a series of modifications in the signaling cascade, including auto-cleavage, cholesterol modification, and palmitoylation, to produce the active form capable of triggering downstream signaling. One of the key players in this pathway is the Shh receptor Patched (Ptc/Ptch) protein. Ptc is a 12-transmembrane protein expressed on the surface of the receiving cell. Shh activates signaling in target cells by binding and inactivating Ptc, which unleashes the 7-transmembrane protein Smoothened (Smo) to stimulate downstream intracellular events and promotes expression of target genes (Alcedo *et al.*, 1996, Denef *et al.*, 2000).

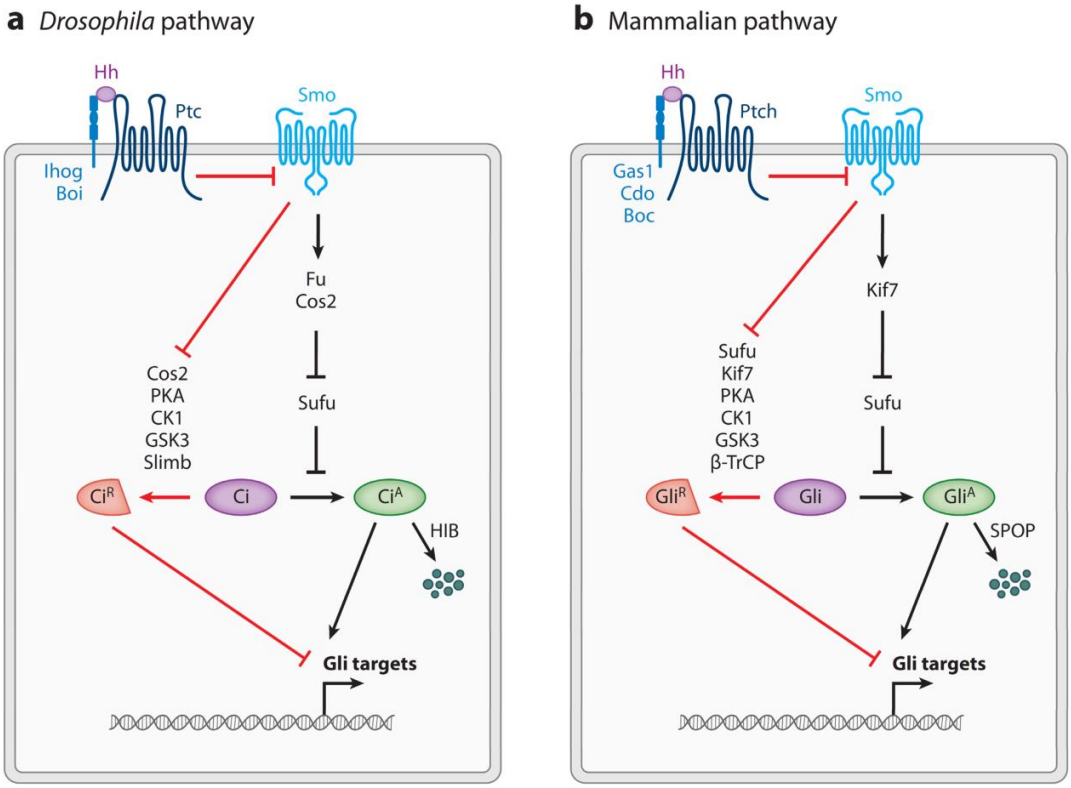

Figure 1. 5 Hh-signaling cascade is highly conserved among insects and vertebrates

Schematic representation of Hh-signaling cascade. a) Drosophila, b) Mammals. When Hh binds with Ptc protein, it releases the inhibition of Smo as a result Smo promotes the formation of a Ci or GLI transcriptional activator (Ci^A or GLI^A) via inhibition of GLI processing and negative upregulation of Sufu and Cos2/Kif7. In the absence of Hh protein Ci or GLI is converted to transcriptional repressor (Ci^R or GLI^R) through limited degradation by proteasome upon phosphorylation by PKA (protein kinase A), GSK3 (glycogen synthase kinase 3β) and CK1 (casein kinase 1) . Adapted and modified from (Hui & Angers, 2011).

Gli transcription factors consist of three family members: Gli1, Gli2, and Gli3, in vertebrates and are the main transducers for Hh signaling. Hh maintains equilibrium among Gli activator (Gli^A) and Gli repressor (Gli^R) activities during development. When Hh is absent, Ptc blocks Smo activity, which allows the formation of Gli^R. Whereas binding of Hh to Ptc stimulates Smo, that prevents Gli^R formation and mediates Gli^A function (Philipp & Caron, 2009).

In *Drosophila*, bifunctional *Cubitius interruptus* (*Ci*) acts as a full length activator (Ci^A) as well as a truncated/processed repressor, when Hh ligand is absent. Full-length *Ci*

is inactivated by inhibitory mechanisms, whereas Ci^R represses a subset of Hh target genes. At both low and elevated levels of Hh, full-length *Ci* exists only as Ci^A. Hh signaling regulates distinct sets of target genes which are switched on or off at different ratios of Ci^R/Ci^A (Methot & Basler, 2001).

In humans, all the GLI family members (*Gli1*, *2* and *3*) have been mapped to separate chromosomal loci by fluorescent *in situ* hybridization (FISH). The first gene to be identified in the GLI family was the human *GLI1* gene (initially called *GLI* and pronounced "glee"), primarily identified by Vogelstein and co-workers as a putative oncogene amplified in glioblastoma cells (Kinzler *et al.*, 1988). GLI1 resides in the 12q13.3 chromosomal region and spans 13 kb of DNA sequence. It includes 11 exons with 3613 bp length, and is translated into a 1106 amino acid protein. Human GLI2 was originally identified as a Tax-helper protein (THP) that binds to Tax-responsive element in the long terminal repeat of the human T-cell leukemia virus (Tanimura *et al.*, 1998). Human *GLI2* resides on 12q14, extending over 257 kb length, having 14 exons, and is translated into a 1586 amino acid product. The last member of GLI family, GLI3, which spans over 260 kb on chromosome 7p14.1 has 15 exons, and is translated into a 1580 amino acid protein product.

Comparing the GLI protein (amino acid) sequences it has been observed that GLI2 is structurally and functionally more similar to GLI3 than GLI1 (92% identity between GLI2/GLI3 and 87% identity between GLI1/GLI2 in their zinc finger domains) (Hui *et al.*, 1994, Ding *et al.*, 1998, Hardcastle *et al.*, 1998). GLI1 is unique in comparison to GLI2 and GLI3 because it only comprises a C-terminal transcriptional activation

domain, while bi-functional GLI2 and GLI3 possess a C-terminal activation and N-terminal repression domain (Figure 1.6) (Dai P, 1999, Sasaki H, 1999).

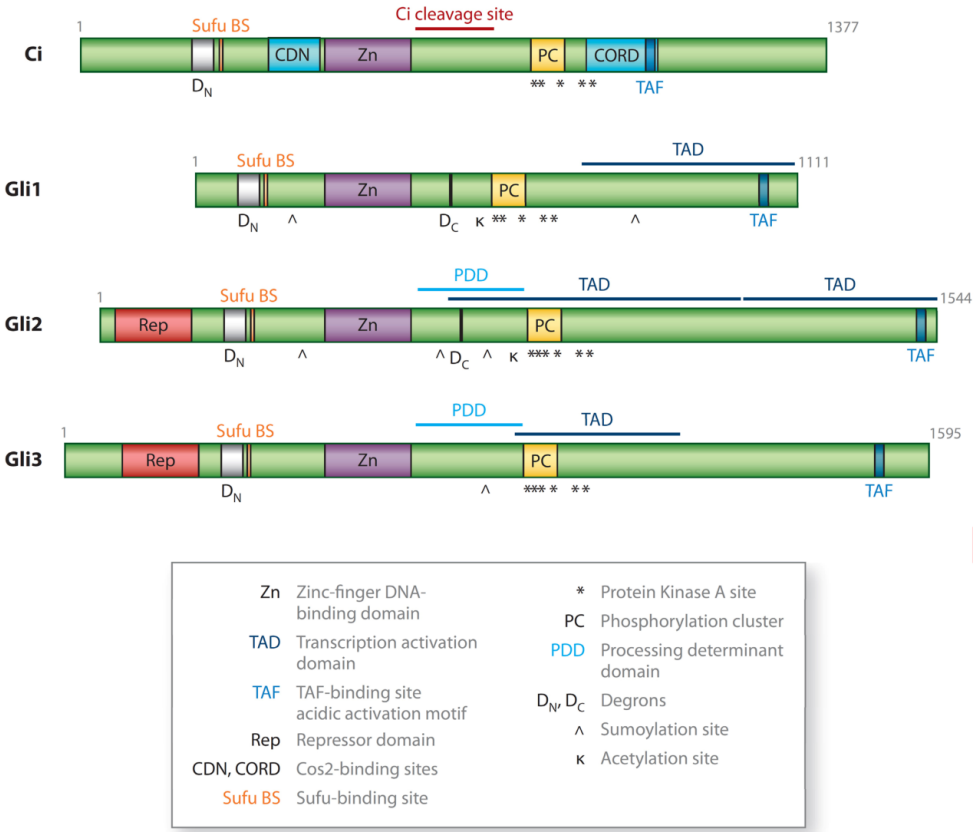

Figure 1. 6 Schematic illustration of domains and motifs in Gli proteins.

Various protein domains and modification sites of the Ci and Gli1, Gli2, and Gli3 are depicted, adapted from (Hui & Angers, 2011).

1.7 Molecular roles of GLI family are highly complex and context dependent

Gli family plays a critical role in mammalian embryonic patterning, more specifically in the central nervous system, the anterior-posterior axis of the embryonic limb bud, craniofacial structures and various internal organs (Ruppert JM, 1998). Gli1 mediates a number of significant cellular processes, such as: migration, invasion, metastasis,

cell proliferation and neural development, through gene regulation. Hh signaling in the epidermis is primarily executed by GLI1. Moreover, GLI1 expresses in the regions close to Shh-expressing cells (Dahmane *et al.*, 1997, Ikram *et al.*, 2004, Hui *et al.*, 1994). Accurate regulation of Hh/GLI signaling is crucial for correct arrangement of the epidermal lineage and development of its derivatives, whereas errors in Hh/GLI signaling disrupts tissue homeostasis and causes basal cell carcinoma (BCC) (Epstein, 2009, Altaba, 1999). GLI1 over expresses in numerous tumors and cancers like glioblastoma, rhabdomyosarcoma, osteosarcomas, BCC and B-cell lymphoma (Roberts *et al.*, 1989, Kinzler *et al.*, 1988, Ghali *et al.*, 1999).

RNA *in situ* studies have shown that GLI1 is expressed in the ventral neural tube and its expression is dependent on Shh signaling (Bai *et al.*, 2002). In knockout mice, Gli1 expression is dependent on Gli2, because GLI1 mutant mice do not have developmental defects or a decrease in Hh signaling, unless one copy of Gli2 is removed (Lee *et al.*, 1997, Matise *et al.*, 1998). However, role of GLI1 protein and other GLI family members is seen to be diverged during vertebrate evolution as zebrafish lacking *gli1* have significant defects in the activation of Hh target genes in neural tube (Karlstrom *et al.*, 2003).

The role of Gli genes in development was first revealed by the discovery of deleterious mutations of GLI3 in several human congenital malformations, including Greig cephalopolysyndactyly syndrome (GCPS) non-syndromic polydactyly, Pallister Hall syndrome (PHS), acrocallosal syndrome, pre-axial polydactyly type IV (PPD-IV) and postaxial polydactyly type A (PAPA) (Elson *et al.*, 2002, Kang *et al.*, 1997, Radhakrishna *et al.*, 1997, Radhakrishna *et al.*, 1999, Vortkamp *et al.*, 1991).

Moreover, GLI3 is also associated with oral-facial-digital syndrome (OFDS) and Opitz syndrome (OS) (Johnston *et al.*, 2010, Liu *et al.*, 2001). A dominant developmental syndrome, GCPS with polydactyly and craniofacial abnormalities, is linked with large deletions, translocations and truncating mutations resulting in functional haploinsufficiency of GLI3 (Johnston *et al.*, 2010, Shin *et al.*, 1999). Mutations affecting murine Gli3, such as extra toe, anterior digit deformity, and polydactyly Nagoya (Pdn), serve as the models for GLI3 morphophathies (Schimmang *et al.*, 1992, Schimmang *et al.*, 1994, Hui & Joyner, 1993, Pohl *et al.*, 1990). The limbs of $Gli3^{-/-}$ mutant mouse embryos show severe polydactyly with many un-patterned digits, A-P polarity is lost and no apoptosis occurs in inter-digital regions (te Welscher *et al.*, 2002)

Multitude studies in mice and other model organisms have suggested that the zinc finger transcription factor GLI3 acts as an antagonist or mediator for the Shh signaling cascade in a context-dependent manner during vertebrate embryogenesis (Coy *et al.*, 2011, Ruppert JM, 1998). It mutually represses its interaction with Shh in the limb and neural tube (Amano *et al.*, 2009, Ruiz i Altaba, 1998). It is an important developmental regulator and is dynamically expressed in the brain, axial, appendicular, and craniofacial structures, as well as within various visceral organs prenatally, postnatally, and in adult life (Lebel *et al.*, 2007, McDermott *et al.*, 2005, Mo *et al.*, 1997, Motoyama J, 1998). Genetic experiments in mice have demonstrated that GLI3 masks the repressor activity of GLI2 (Bowers *et al.*, 2012). However, Gli3 acts as an activator in the absence of GLI2 (McDermott *et al.*, 2005). During neural tube formation, spatio temporal expression of Gli3 is crucial to ensure

accurate Hh signaling. At the start of neurulation, GLI3 is expressed throughout the neural tube and is later on restricted to the dorsal neural tube (Alvarez-Medina *et al.*, 2008).

1.7.1 Gli2 is critical in Hh signaling for development of neural tube, limb and internal organs

Gli2 is a bifunctional transcription factor containing both activator and repressor domains and is involved in a multitude of patterning mechanisms during early vertebrate development (Qi *et al.*, 2003). Several studies in mice show that the zinc finger protein Gli2 is widely expressed in the neural tube and mainly acts as an activator in the two most ventral domains, i.e floor plate (FP) and pV3 interneurons (McDermott *et al.*, 2005, Bai & Joyner, 2001, Park *et al.*, 2000, Lee *et al.*, 1997). Diverse ventral patterning defects in hindbrain and spinal cord are seen in *Gli2−/−* mice, with a severely affected floor plate (FP) and interneurons, and thus there is no survival until birth. These defects are more severe than *Gli3−/−* mutants (Ding *et al.*, 1998, Lebel *et al.*, 2007). In the spinal cord, *Gli2* is necessary for oligodendrogenesis and *Gli2* mutation leads to a delay and decrease in the production of oligodendrocytes (Qi *et al.*, 2003). Interestingly, Gli2 co-expresses with Shh in the developing notochord, but Gli2 mutant mice have a normal notochord (Ding *et al.*, 1998). Loss-of-functional studies in mice demonstrate that *Gli2* is associated with craniofacial abnormalities, microcephaly, and flattened head showing subcutaneous odema (Mo *et al.*, 1997). Consistent with genetic studies in mice, patients with *GLI2* mutations have been associated with holoprosencephaly, branchial arch anomalies and CNS abnormalities (Rahimov *et al.*, 2006). Furthermore, Gli2 has also been

implicated in skeletal muscle formation, is initially expressed throughout the somites, and is then predominantly found in the dorsal somite (McDermott *et al.*, 2005). Overexpression of activator function of Gli2 in transgenic mice induces disruption of primary cilia resulting in basal cell carcinoma and medulloblastoma (Mager *et al.*, 2008). In chicken, Gli2 is expressed throughout the neural plate, between the rhombomeres and dorsal dermomyotome (Paquet *et al.*, 2009).

In contrast to mammals, several teleost species have an additional copy of *Gli2* gene (*gli2a/gli2b*), due to an extra round of whole genome duplication in teleost, introducing more complexity to Hh signaling. In zebrafish, *gli2a*/*gli2b* plays an activating role during cell proliferation and differentiation in the sensory neurons, branchial motor neurons, FP and hindbrain (Ke *et al.*, 2008, Anelli *et al.*, 2009). RNA *in situ* studies in zebrafish detected *gli2a* in the anterior neural plate and as the development proceeds *gli2a* can be detected uniformly throughout the dorsal forebrain, midbrain, hindbrain, and is also expressed in the adjacent and dorsal to those cells expressing *Shh* (Karlstrom *et al.*, 1999, Ding *et al.*, 1998, Ruiz i Altaba, 1998). Several studies have shown that *gli2a* is preferentially expressed in the lateral mesoderm, whereas *gli2b* is expressed predominantly in the neural tube. Moreover, the combined activity of *gli2a*/*gli2b* is a prerequisite in hedgehog signaling for neuronal development (Ke *et al.*, 2008). In zebrafish, most of the Gli2 functions are conserved, however *gli1* rather than *gli2* appears to be the major activator of the early Hh response in the CNS (Distel *et al.*, 2009). In contrast to mouse, zebrafish that lacks *gli2* have only negligible abnormalities in Shh signaling, strengthening the

idea of functional divergence of Gli proteins during vertebrate evolution (Abbasi *et al.*, 2009).

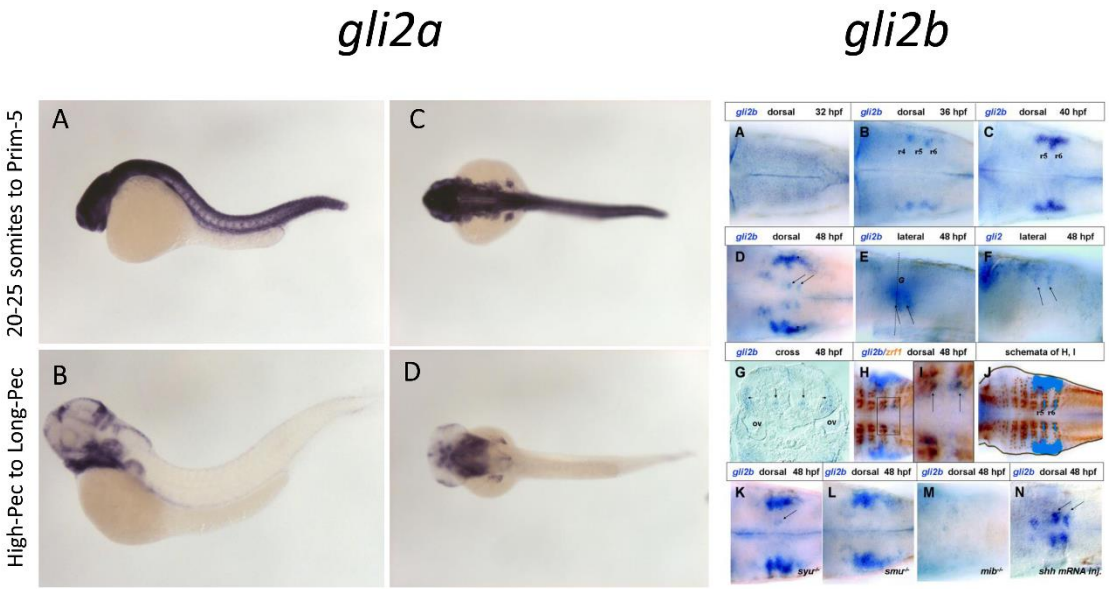

Figure 1. 7 RNA in situ hybridization studies showing gli2a and gli2b expression in several embryonic domains at 24 and 48hpf *(Source: www.zfin.org)*

Gli2 is essential for normal endochondral bone development and has thus also been implicated in limb development. The anterior part of the limb bud expresses Gli2, therefore Gli2 seems to prepattern the embryonic limb bud along its anterior posterior (A/P) axis (Theil *et al.*, 1999, Bowers *et al.*, 2012). Similarly, in chicken, Gli2/Gli3 are initially expressed throughout the limb bud mesenchyme but later on the expression is downregulated in posterior mesenchymal cells (Paquet *et al.*, 2009). Gli2 co-expresses with Gli3 and regulates digit identity in the anterior-posterior regions of autopod (Bowers *et al.*, 2012). $Gli2^{-/-}$ mutant mice show a significant reduction in the stylo- and zeugopod of the fore and hindlimbs (Mo *et al.*, 1997). Gli2 mutations are associated in patients with facial midline anomalies, pituitary defects, preaxial and postaxial polydactyly (Bertolacini *et al.*, 2012).

Mutations in human *GLI2* have been linked with holoprosencephaly, pituitary anomalies (hypopituitarism), congenital growth hormone deficiency, cranial and midline facial deformities, and abnormalities in limb development (pre-axial and post-axial polydactyly) (Bertolacini *et al.*, 2012, Franca *et al.*, 2013, Roessler *et al.*, 2003, Roessler *et al.*, 2005). *Gli2* also plays a crucial role in carcinogenesis and transgenic mice studies suggest that over-expressing *Gli2* in cutaneous keratinocytes develops multiple BCC, while the human hepatocellular carcinoma cell lines and hepatocellular tissues show high levels of *GLI2* (Grachtchouk M, 2000, Cheng *et al.*, 2009, Tojo *et al.*, 2003). Other human tumors associated with *GLI2* are medulloblastomas, prostate and breast cancer (Fulda *et al.*, 2002, Okano *et al.*, 2003, Sicklick *et al.*, 2006). The widespread expression of Gli2 provides compelling evidence for its dynamic role in Shh signaling during embryogenesis.

1.8 *Cis*-regulatory underpinning of GLI family

As *GLI* family is very critical in Shh signaling and crucial for patterning of many aspects of the vertebrate body plan, studying the regulation of these key genes is essential to understand the developmental pathways and pathological conditions in detail. Characterization of the promoter region and genomic organization of GLI1 has already been done (Liu *et al.*, 1998). Similarly, the *cis*-acting regulatory catalogue of human *GLI3* gene was reported recently (Paparidis *et al.*, 2007, Abbasi *et al.*, 2007, Abbasi *et al.*, 2010, Abbasi *et al.*, 2013). Twelve intronic human-fugu conserved noncoding elements (CNEs) from the introns of *GLI3* were shown to operate in transiently transfected cultured cells in a cell-type dependent fashion, as activators or repressors of reporter gene expression. The activating or repressive potential of

the CNEs observed in human cell culture was retained *in vivo* in zebrafish and mice embryos. The expression induced by these elements was in agreement with already reported *gli3/Gli3* expression in zebrafish and mice (Abbasi *et al.*, 2007, Abbasi *et al.*, 2010, Abbasi *et al.*, 2013). *GLI2* has been studied most extensively during neural tube and limb development. However, the genetic mechanisms and the *cis*-acting elements controlling the expression of *GLI2* remains largely unknown.

1.9 Aims and objectives

Complex spatiotemporal and quantitative aspects of GLI2 expression signal the occurrence of a highly sophisticated network of *cis*-acting regulatory catalog to orchestrate the partitioning of its activity domains for the correct interpretation of Hh signaling cascade. Towards the elucidation of the basic regulatory network of signaling molecules and the associated transcriptional regulators patterning the vertebrate body, this thesis will focus on the detection and functional analysis of *cis*-acting regulatory elements of the key developmental regulator GLI2, using zebrafish as a model.

In this study, in an attempt to define and characterize the *cis*-acting regulatory catalogue of *GLI2,* the following steps will be undertaken:

- Multi-species (tetrapod/teleost) comparative sequence analysis will be done to identify *cis*-acting regulatory modules of *GLI2*,
- Validation of these conserved intervals by transient reporter gene assay in zebrafish embryos using co-injection,

- Verification of *GLI2*-specific gene regulatory intervals by *Tol2* vector based transgenesis, and
- Functionally active modules will then be further tested for TFBSs by various phylogenetic footprinting tools.

MATERIALS AND METHODS

2.1 *In silico* protocols

2.1.1 Sequence collection and comparative analysis

Human *GLI2* genomic sequence (ENSG00000074047) was obtained from Ensembl genome browser (http://www.ensembl.org) release65 (GRCh37) with 100kb flanking region along with orthologous sequences from mouse (NCBIM37), chicken (Galgal4), fugu (Fugu4) and zebrafish (Zv9). The annotation file containing information about protein coding and non-protein coding regions of human *GLI2* gene was also retrieved. Multi-species sequence comparison was performed using the MLAGAN alignment tool and visualized by VISTA (mVISTA) (http://genome.lbl.gov). Human sequence was used as a baseline and annotated by using exon/intron information available at Ensembl genome browser (Brudno *et al.*, 2003). Human sequence was masked for human/primates to get a better picture of the alignment (Mayor *et al.*, 2000). Conservation was measured using a 50bp window and a cutoff score of 50% identity.

2.1.2 *In silico* mapping of conserved TFBSs within each CNE

To identify putative conserved transcription factor binding sites (TFBSs) for each CNE, the orthologous sequences of terrestrial and non-terrestrial vertebrates were retrieved from Ensembl genome database by BLAST-N (Basic Local Alignment Search Tool for Nucleotide) based similarity search. Each of the *GLI2*-associated CNE and its orthologous sequences were analyzed using MEME (Multiple Expectation Maximization for Motif Elicitation) motif discovery algorithm (Bailey *et al.*, 2009). MEME is a PWM (position weight matrixes) based algorithm, which identifies over-

represented motifs in the query set. Considering the expected length of transcription factors binding sites, the criteria for minimum length was set from 6 to 12 bp. The identified motifs of each CNE were characterized further by using the STAMP tool to determine known transcription factors against TRANSFAC (TRANScription FACtor) v11.3 database (Matys *et al.*, 2003, Mahony & Benos, 2007). Each of the specified transcription factors were then screened for endogenous gene expression studies (RNA *in-situ* hybridization) using MGI (Mouse Genome Informatics) database (http://www.informatics.jax.org/).

2.1.3 Syntenic analysis

In order to associate each of the selected subset of identified CNE with their probable target gene (*GLI2*), each CNE was analyzed using BLAST and a gene synteny was drawn using BLAST (Ensemble and UCSC genome browser) for *GLI2* gene and flanking genes, between tetrapod and teleost. This allowed us to map carefully the genomic context of evolutionary conserved elements in corresponding fugu and zebrafish loci. Among these anciently diverged genomes (human-teleost fish, >450 Mya) uninterrupted physical linkage between CNEs and *GLI2* gene was taken as an evidence of functional association.

2.2 DNA protocols

2.2.1 Genomic DNA extraction from whole blood

Genomic DNA was extracted from whole blood using standard phenol-chloroform procedure. Five hundred microliters of human blood was taken in an eppendorf tube along with 750µl of solution A, mixed by inverting the tubes 4-6 times and kept at room temperature for 15-30 minutes. The tube was centrifuged at 13,000 rpm for 1 minute and the supernatant was discarded. The nuclear pellet was re-suspended in 400µl of solution A. The tube was again centrifuged for 1 minute at 13,000 rpm and after discarding the supernatant the nuclear pellet was re-suspended in 400µl solution B, 12µl of 20% SDS and 10µl of proteinase K (10 mg/ml). The sample was incubated at 37^{o}C overnight. On the following day, when the pellet was completely digested, 500µl of a fresh mixture of equal volume of solution C and solution D was added in the sample, mixed and centrifuged for 10 minutes at 13,000 rpm. The aqueous upper layer was collected in a new tube and equal quantity (500µl) of solution D was added and centrifuged at 13,000 rpm for 10 minutes. Again the aqueous upper phase was transferred to a new tube and DNA was precipitated by adding 55µl of sodium acetate (3M, pH 6) and equal volume (500µl) of isopropanol (stored at -20˚C). The tubes were inverted several times gently to precipitate the DNA and centrifuged at 13,000 rpm for 10 minutes. The supernatant was then carefully discarded without disturbing the DNA pellet. To the DNA pellet obtained, 200µl of 70% ethanol (stored at -20˚C) was added and centrifuged for 7 minutes at 13,000 rpm. The ethanol was then discarded and the DNA dried by keeping the tubes for 10 minutes at room temperature. The precipitated DNA was then dissolved in an

appropriate amount of Tris-EDTA (TE) buffer. The samples were then kept in an incubator for a day to ensure complete suspension of the DNA in the buffer.

The genomic DNA, PCR purified products and plasmids were quantified by NanoDrop 1000 spectrophotometer (Thermo Scientific) at 260nm wavelength. A 260:280 absorbance ratio of ~1.9 for pure DNA and ~2.00 for RNA was expected. A ratio of A_{260}/A_{280} was used to determine the purity of the DNA and a ratio between 1.9 and 2.1 represented a high-quality DNA sample. The DNA outside the range of 1.9 to 2.1 was either discarded or re-purified using ethanol precipitation method.

2.2.2 Ethanol precipitation to purify the DNA

The reaction was carried out in a 0.2ml eppendorf tube. Total volume of the DNA sample was measured through pipette, one-tenth volume of 0.3M sodium acetate (pH 5.2) was added, and mixed well. Two volumes of 100% ice-chilled ethanol were then added and mixed well. The reaction tube was placed at -20°C for 30 minutes. The reaction mixture was centrifuged at 14000 rpm for 15 minutes and the supernatant was discarded. One ml 70% ethanol was added and mixed gently followed by a spin at 14000 rpm for 5 minutes. Supernatant was again discarded and the pellet was diluted in 30µl volume of molecular-grade water or TE.

2.3 PCR Protocols

2.3.1 Primer designing and dilution

PCR primers were designed using manual analysis and two different algorithms available at Primer3 (http://gmdd.shgmo.org/primer3) and sequence manipulation suite (http://www.bioinformatics.org/sms2/). Primers were diluted in molecular-

grade water according to instructions provided by the manufacturer. Specificity of the primers was verified using UCSC (http://genome.ucsc.edu/) and NCBI (http://www.ncbi.nlm.nih.gov/) browsers. The details of the primer sequences used to amplify the CNEs are given in Table 2.1.

2.3.2 Polymerase chain reaction (PCR)

PCR amplification was carried out in 0.2ml tubes, according to a standard procedure, in a total volume of 25µl containing: 1µl DNA dilution, 1µl each of forward and reverse primer (20 ng/µl), 2.5µl 10X PCR buffer (200 mM $(NH_4)_2SO_4$, 750 mM of Tris-HCl pH 8.8), 1.5µl 25 mM $MgCl_2$, 0.5µl 10 mM dNTPs and 0.2µl of 0.5 unit *Taq* DNA Polymerase (Thermo Scientific) in 17.3µl PCR water. The reaction products were centrifuged for 30 seconds at 8,000 rpm for thorough mixing. Reactions were performed by means of PTC-200 DNA Engine Cycler (Bio-Rad, USA). The reaction mixture was taken through cycling conditions consisting of: 5 minutes of an initial DNA denaturation at 95°C, followed by 35 cycles of amplification, each consisting of 3 steps: 30 seconds at 95°C for denaturation of DNA into single strand, 30 seconds for annealing at a temperature at which the primer is to hybridize or anneal, and 1 minute at 55 to 60°C for extension of complementary DNA strands from the primers.

This was followed by a final 10 minutes at 72°C for *Taq* DNA polymerase to synthesize any unextended strand left. PCR products were analyzed on 2% agarose gel, as described above. After analysis on an agarose gel, PCR products were then purified using PureLink PCR Purification Kit (Life Technologies). Purified PCR product (1µl) was verified on a 2% agarose gel.

Table 2. 1 Primers used to amplify the intra-*GLI2* conserved non-coding elements (CNEs) for co-injection assay and cloning into *Tol2* vector

Element	Forward primer	Reverse primer	Annealing temperature
CNE1	CCGATGACTGAAGCCATAGC	CTGGTAAGGAGGTGGAGCAC	61˚C
CNE2	CGGCTCCACACTATCCTCAAG	GAATGAGAGAGTGCAGGGAACAC	60˚C
CNE3	CTGCTTAGATGGCACCTTGC	GGAGTGAAGGGCAGCAAGTC	60˚C
CNE4	GCCCACACACCCAGCCTAGC	GCTGCAAAGACCTCTTCCGAG	60˚C
CNE5	CAACTTCCCAAGTGACTGTGTT	CTAGAAGGCAGAGGCAACGTC	60˚C
CNE2_Short	GCTCTGGCTTCCATGATGAATG	CACGAAATGGTCACCAGCTCAG	59˚C
Dr-*gli2*_CNEa	CTGTCTCCTTTGATTAATGTGAC	GGAACACCTATGCATGTACCA	58˚C
Dr-*gli2*_CNEb	GCCTGTCTCAACGCTTGATTA	AGCTCCGAGACCCCTGACAT	58˚C

2.3.3 DNA purification with gel extraction kit

PureLink quick gel extraction kit (Life Technologies) was used for the purification of the PCR purified products, according to the following protocol: The total volume of PCR product was run at 90V on a 2% agarose gel for ~25 minutes and sliced from the gel. The sliced piece of gel was placed in a 2ml eppendorf tube. The gel slice was weighed, and 4 volume of gel solubilisation buffer (provided in the kit) was added. The eppendorf tube was then incubated at 50°C for 10-15 minutes accompanied by mixing after every 3 minutes. After the gel was completely dissolved, the solution was transferred to a column provided with the kit, and centrifuged at >12,000rpm for 1 minute. The flow-through was discarded and the column was washed with 500µl of wash buffer provided with the kit by centrifuging at >12,000rpm for 1 minute.

After drying the column it was placed into a fresh 2ml eppendorf tube. To elute DNA, 25-40µl of HPLC water was applied to the center of the column and centrifuged at

>12,000rpm for 1 minute. The eluted products were then analyzed on a 2% agarose gel and quantified by NanoDrop (Thermo Scientific), as described above.

2.4 Bacterial Protocols

2.4.1 Preparation of media and agar plates

LB (Luria-Bertani) broth was prepared by adding 10g of bactotryptone, 5g of yeast extract and 10g of NaCl to 1L of boiling double-distilled water (ddH_2O) and mixing until homogeneous. The pH was adjusted to 7.5 by adding 2ml of 2N NaOH. LB agar plates were prepared by adding 15g of agar to 1L LB media and sterilized by autoclaving at 121^oC for 15 minutes. Media was then stored at 4^oC, until needed.

2.4.2 Preparing competent cells and storage

Escherichia coli DH5α cells were grown overnight for 18 hours on non-antibiotic LB agar plates at 37°C. One colony was inoculated into 5 ml LB broth at 37°C and grown overnight for 18 hours. Five hundred microliters of the overnight culture was inoculated into 100 ml LB broth. The culture was grown with 250rpm at 37°C until it reached A_{600} of 0.6. The culture was then incubated on ice for 30 minutes, prior to centrifugation for 10 minutes at 4000rpm at 4°C. Cell pellets were then suspended in 5ml ice cold 0.1M $CaCl_2$ and further incubated on ice for 10 minutes. The cell pellets were then centrifuged and suspended again in 1ml ice cold 0.1M $CaCl_2$ and placed on ice for 1hour and centrifuged for 10 minutes at 4000rpm at 4°C. The supernatant was discarded and the pellet was suspended in 4:1 $CaCl_2$: glycerol mixture and the cells were then aliquoted in eppendorf tubes for storage at -80°C.

To store a single colony was picked from a plate and grown for 18 hours in 5ml LB broth. Half a milliliter of the overnight culture was added to 0.5ml of the 80% sterile glycerol in the eppendorf tube. The tube was vortexed and the glycerol stocks were stored at -80°C freezer. To refresh a culture from a glycerol stock, eppendorf tubes were taken from -80°C freezer and 50µl of stored culture from the top of the frozen glycerol stock and was streaked onto a pre-prepared LB agar plate. The plate was incubated for 18 hours overnight at 37°C.

2.4.3 Transformations

Chemically competent DH5α and commercially available TOP10 cells were used for transformation. Two microliter from the TOPO or ligated (CNE and vector) reaction were added to the competent cells and incubated on ice for 30 minutes. Heat shock was given to the cells for 30 seconds at 42°C to change the fluidity of the *E.coli* cell membrane. DNA could then enter into the bacterial cell at an efficient rate by cell surface invagination. The cells were immediately transferred on ice and 250µl of LB media was added. The reaction was incubated at 37°C with a shaking speed of 250rpm for 1.5 hour. Pre-warmed LB agar plates, with spectinomycin (100µg/ml) for TOPO constructs or ampicillin (100µg/ml) for *Tol2* plasmids, were used to streak with 75µl of freshly grown culture. The plates were incubated for 18 hours at 37°C. After 18 hours incubation, the selected colonies were transferred aseptically to 15ml falcon tubes containing 3ml LB media with spectinomycin/ampicillin antibiotic (100µg/ml). The falcon tubes were incubated for 18 hours in shaking incubator (250rpm) at 37°C. After completion of incubation period, 1.5ml from each falcon tube was proceeded for bacterial plasmid isolation.

2.4.4 Plasmid preps

Plasmid was purified by using the PureLink Quick Plasmid MiniPrep kit (Life Technologies) according to manufacturer's instructions. The purified plasmid was run on a 1% agarose gel and was quantified by NanoDrop 1000 Spectrophotometer (Thermo Scientific), as described above.

2.5 Cloning CNEs into the transposon vector

This method was adopted from Fisher et al. (2006). The destination vector into which identified CNEs were cloned is pGW_*cfos*EGFP (Fisher *et al.*, 2006).

2.5.1 Generation of entry clone

pCR8/GW/TOPO® TA vector cloning kit (Life Technologies) was used to generate the entry clone. CNEs were amplified from genomic DNA of human and zebrafish using the standard PCR procedure, as described above. However, the final extension step at 72°C was extended to 20 minutes to produce 3'-adenine overhangs to each end of the PCR product. Fresh PCR products of CNEs were analyzed by loading 5µl of the product on 2% agarose gel, and the reaction was performed in an eppendorf tube, according to the standard procedure: ~2µl freshly amplified PCR product, 0.5µl pCR8/GW/TOPO TA vector (see plasmid map in appendix, Figure A1), 1µl salt solution (1.2M NaCl 0.06M $MgCl_2$), and 6µl molecular-grade water. The reaction was mixed gently and incubated at room temperature for 5 minutes. After 5 minutes, the TOPO reaction was shifted on ice to proceed for transformation.

2.5.2 Restriction digestion of TOPO entry clone with *EcoRI*

As pCR8/GW/TOPO® TA vector has two restriction sites of *EcoRI* flanking the TOPO. These sites can be used to confirm the insertion of desired fragment (i.e. the CNE). Purified plasmid was digested to excise inserted CNEs from TOPO entry clone plasmid. Standard procedure was followed for total volume of 10µl containing 1µl 10X React 3 Buffer (Life Technologies), 0.25µl *EcoRI* (Life Technologies), 8µl plasmid (concentration=100-150 ng/µl) and the volume was made up to 10µl by the addition of ddH_2O. The digestion mix was incubated at 37°C in a water bath for 90 minutes. The whole digestion mix was loaded on a 2% agarose gel, and the gel was run on 90 V for 35 minutes. DNA ladder of 100 bp (Thermo Scientific) was loaded in parallel to confirm the size of the insert.

2.5.3 Orientation screening of insert-TOPO vector clones

Once the insertion of CNEs was confirmed by *EcoRI* digestion, the correct orientation (sense strand) of an insert (CNEs) into TOPO vector needed to be identified. TOPO entry clone was either sequenced commercially (Macrogen) or screened using combination of TOPO-specific and CNEs-specific forward and reverse primer pairs. The sequenced results were analyzed by Bioedit sequence alignment editor version 7.0.9.0.

2.5.4 Generation of destination clone or recombination with pGW_*cfos*EGFP destination vector

LR recombination reaction was done between ~100 ng/µl TOPO entry clone and ~100 ng/µl destination vector pGW_*cfos*EGFP (see plasmid map in Appendix, Figure A2). Total volume of 10µl for LR reaction was prepared containing 1µl

TOPO vector with insert (100ng/µl), 1µl pGW_*cfos*EGFP (~100ng/µl), 6µl TE buffer and 2µl Gateway LR clonase II enzyme (Life Technologies). LR reaction mix was incubated at room temperature for 60 minutes. The activity of LR clonase was stopped by the addition of 1µl proteinase K and further incubated at 37°C for 10 minutes. Two µl of LR mix was used to transform TOP10 competent cells, as described above. Plasmids were purified using the PureLink Quick Plasmid MiniPrep kit (Life Technologies) as per manufacturer's instructions, and eluted in 50µl high grade molecular water. The purified plasmid was sequenced with *Tol2* specific primers for checking orientation and validation of the insert into destination vector.

2.6 *In vitro* transcription of transposase RNA

This method is based on that of Fisher et al. (2006). pCS-TP vector was linearized using standard digestion protocols, as described previously (see plasmid map in Appendix, Figure A3). Linearized plasmid DNA and PCR products that contain an RNA polymerase promoter site can be used as templates for *in vitro* transcription. About 2µg of pCS-TP vector was taken into an eppendorf tube along with 0.5µl of NotI, 5µl of buffer O (Thermo Scientific) and 26µl of high molecular-grade water. The tube was mixed and centrifuged for 5-10 seconds and kept at room temperature for 2 hours. The digestion was verified on a 2% agarose gel by loading 1µl digested and undigested products in parallel wells. Once the digestion was verified, all the samples were loaded on a 2% agarose gel and purified by PureLink Gel Purification kit (Life Technologies), and eluted in 14µl of TE buffer provided with the kit. About

1µl of the purified product was verified on 2% agarose gel and concentration measured using NanoDrop (Thermo Scientific).

2.6.1 Capped transcription reaction assembly

Six microliters of the linearized vector (19 ng/µl) was taken into fresh eppendorf tube along with 10µl of NTP/CAP mix, 2µl of SP6 polymerase, 2µl of SP6 buffer (10X) and incubated at room temperature for 2.5 hours. After the incubation period, 1µl of TURBO DNAse was added to remove the template DNA, mixed well and re-incubated at room temperature for 15 minutes.

2.6.2 Recovery of RNA by lithium chloride precipitation

Reaction was stopped and precipitated by adding 30µl nuclease-free water and 30µl of lithium chloride precipitation solutions and kept at -20°C for 1.5 hours. The solution was centrifuged at 4°C for 30minutes at maximum speed to pellet the RNA. The supernatant was carefully removed and the pellet was washed with 1ml of ice chilled 70% ethanol. The solution was re-centrifuged at maximum speed to remove the unincorporated nucleotides. The ethanol was then discarded and the RNA dried by keeping the tube for 10-20 minutes at room temperature. The RNA was then re-suspended in RNAse free water provided with the kit. The concentration of RNA was determined by NanoDrop and length of the product was verified on a 2% agarose gel, i.e. ~900bp.

2.7 Generation of transgenic zebrafish

2.7.1 Zebrafish breeding and mating

Zebrafish were bred and raised according to standard protocols. The light and dark cycle (14 hours light and 10 hours dark) was controlled by an automatic timer, whereas temperature was controlled by an automated electric heater at 28.5°C. Water pH was maintained at approximately 7.0 and quality was controlled by the combination of three different types of biological filters (activated carbon, ammonia remover and ceramic rings). Fish were well fed with rich source of proteins and carbohydrates like brine shrimp and flakes.

Female zebrafish laid eggs when lights were turned on in the morning. To set up for mating, equal number of healthy and active males and females were selected in separate tanks for breeding. The males and females were kept separate by placing a divider in the tanks overnight. Divider was removed next day in the morning when light was turned on and fish were free for mating. After 25-40 minutes, 1-2 cell stage embryos were collected, cleaned and washed with distilled water.

Transient transgenic zebrafish was generated using two independent reporter assays, i.e. co-injection and *Tol2* based reporter system with the help of micro glass capillaries and Femtojet pressure injection system (Eppendorf). Zebrafish strains used for the transgenic assays are: Queen Mary wild type, Tubingen wild type and rh3/5:KalTA4 (aka: r3r5 RFP).

2.7.2 Needle pulling and ramp test

P-97 Flaming/Brown micropipette puller (Sutter instrument company, USA) was used to pull the Kwik-Fil borosilicate glass capillaries (World Precision instruments) using the following micropipette program parameters: PULL 60, VELOCITY60 and TIME 20ms. Heat was adjusted after testing each lot of the capillaries using a ramp test.

The purpose of the ramp test is to determine the heat value required to melt the glass capillaries. This test should be run before the start of a new heating filament, a new lot of glass capillaries, or before writing or editing a program, according to the following protocol: a desired program was selected from 1 to 99 by pressing the enter button, followed by pressing clear button. 0 was pressed to avoid deletion of a parameter. Option 1 was pressed for ramp test. The lid was opened and glass capillary was installed into the puller bars, and the pull button was pressed. After completion of ramp test, the heat value was shown on the main display required to melt the particular lot of glass capillaries. The newly determined valued (around 500) was entered into the desired program.

2.7.3 Co-injection reporter assay

This assay is adapted from a protocol presented by (Muller *et al.*, 1999). Genomic DNA of human and zebrafish was used to amplify (standard PCR procedure as described above) the conserved non-coding regions for transient transgenic zebrafish co-injection assays. The reporter expression cassette consisting of EGFP under the control of mouse β-globin minimal promoter was amplified from plasmid vector (see plasmid map in Appendix, Figure A3) by PCR (Thermo Scientific DNA Taq) and purified using the PureLink PCR purification kit according to manufacturer's

instructions (Life Technologies). PCR purified product of CNEs (30ng/µl) and β-globin-GFP promoter-reporter cassette (15ng/µl) were combined, and 0.5% phenol red (Sigma) was used as a tracer dye, as described previously (Woolfe *et al.*, 2005). The injection mix was injected with pre-prepared glass or pulled microcapillaries, using Femtojet pressure injection system (Eppendorf) at the 1 to 2-cell stage of embryonic development, under a stereo microscope CSM2 (Labomed).

2.7.4 *Tol2* mediated transgenesis

The destination clones consisted of CNEs and minimal *c-fos* promoter were sequenced for conformation of positive orientation into transposon construct. The purified transposon construct (25ng/µl), 0.5µl transposase RNA enzyme (175ng/µl), and 0.5µl phenol red stock, were injected into one-cell stage zebrafish embryos. The transient transgenic embryos were screened at day2 and day3 using a Leica MZ 16F, IX81 or IX71 fluorescence stereomicroscope, and photographs taken with a Leica DFC310FX or Olympus DP72 camera.

2.7.5 Post injection treatments and embryo screening

Embryos developing abnormally were discarded after 4 to 5 hours of injection. The injected embryos were raised at 28.5°C in 1X embryo medium containing 0.003% PTU (phenylthiourea) to prevent pigmentation. The zebrafish embryos were dechorionated manually by fine forceps at day2 and anaesthetized by Tricaine. The transgenic embryos were screened for GFP signals under an inverted fluorescent microscope (IX71, Olympus) using DP72 microscope digital camera or DP2-TWAIN (Olympus). GFP expressing domains were classified according to the following tissue categories: forebrain, midbrain, hindbrain, spinal cord, eye, ear, notochord, muscle,

circulating blood or blood islands, heart or pericardial region, epidermis, and fins. GFP expressing cells that were not localized unequivocally were classified as “others”. The schematic for location and tissue-specific expression of each CNE was generated using Adobe Photoshop software.

RESULTS

3.1 Identification of tetrapod-teleost conserved non-coding elements in *GLI2* locus by comparative sequence analysis

A multi-species alignment of human *GLI2* with orthologous genomic sequences from mouse, chicken, fugu, and zebrafish, revealed five anciently conserved non-coding elements embedded exclusively in the intronic intervals of GLI2 gene (Figure 3.1), maintaining at least 50% identity over a 50bp window across all species. A stringent criteria of at least 50% conservation over a 50bp window across all species, was employed to highlight the tetrapod-teleost conserved regions, as shown by previous studies that a criteria of 50-70% is ideal to identify putative regulatory elements (Abbasi *et al.*, 2007, McEwen *et al.*, 2006, Woolfe *et al.*, 2005). Comparative analysis of 100kb upstream or downstream region of the GLI2 gene did not detect any significant sequence similarity from human to fish, and all the CNEs were located within the introns. CNE1 and CNE5 are present within intron 1, while highly conserved CNE2 is present within intron 2, CNE3 is located in intron 4, and CNE4 is positioned within intron 7. Sequence comparison of human *GLI2* with other tetrapods like mice and chicken shows large number of conservation peaks due to short evolutionary distance, whereas human and teleost sequence comparisons the number of CNEs drops sharply (Figure 3.1). A subset of intra-*GLI2* genomic intervals conserved down to fish were prioritized for functional assays. The details of conserved tetrapod/teleost amplicons selected for functional analysis are described in Table 3.1. Further analysis of the selected CNEs was done using the UCSC and

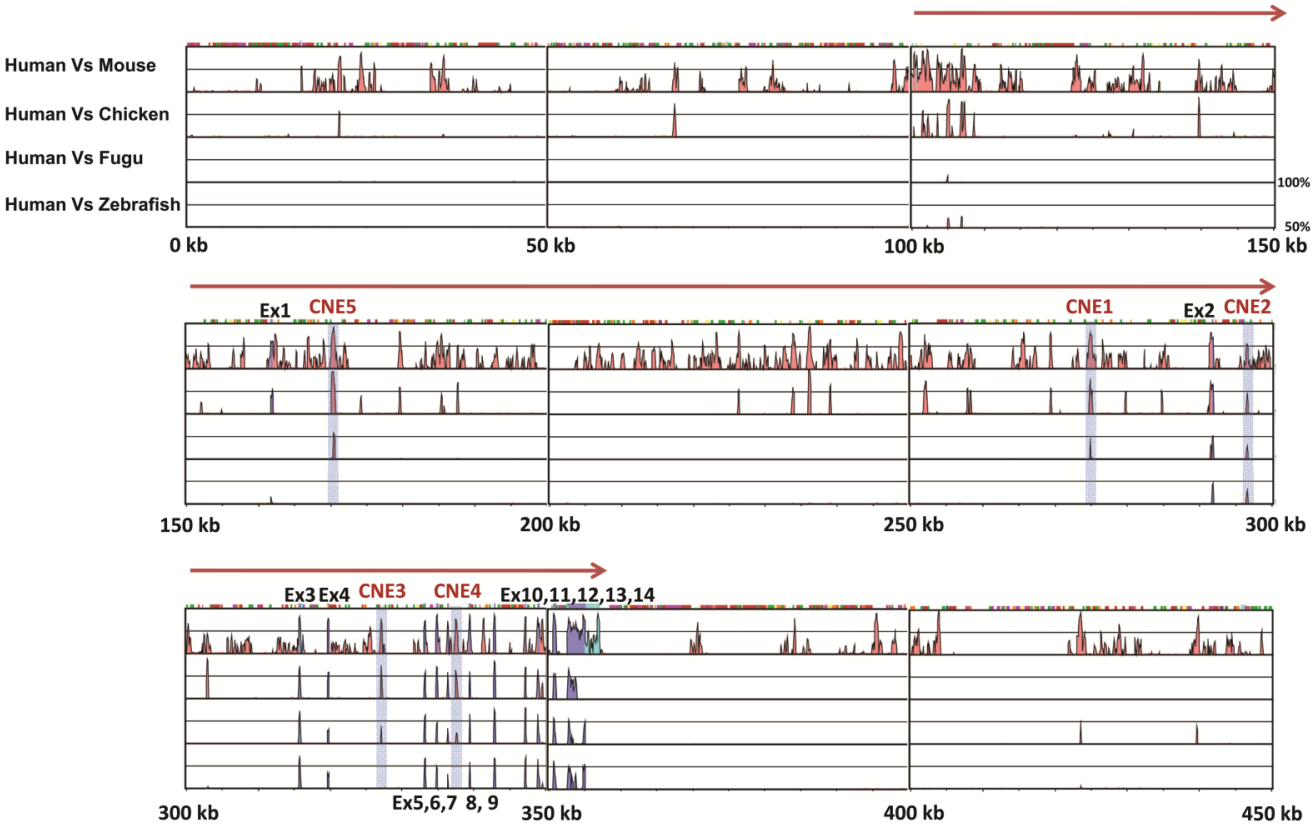

Figure 3. 1 MLAGAN alignment of the genomic region encompassing *GLI2*

In each panel, human genomic GLI2 DNA sequence obtained from ENSEMBL and is aligned with mouse, chicken, fugu and zebrafish orthologous regions. Alignment parameters are explained in the methods section.CNE1-CNE5 that have been selected for functional assay are color shaded. Ex and CNE stand for exon and conserved non-coding element, respectively. Conserved coding and non-coding sequences are depicted in blue and pink respectively.

Table 3. 1 Tetrapod-teleost conserved non-coding elements (CNEs) from Intron of human *GLI2* selected for functional analysis in transgenic zebrafish assay

Element	Region	Amplicon Coordinates Chr2	Amplicon Size (bp)	Conservation Human-Fugu 50%; >50 bp
CNE1	Intron 1	121667550–121668542	993	72% (131bp)
CNE2	Intron 2	121689507–121690037	531	67% (128bp)
CNE3	Intron 4	121719881–121720730	850	74% (73bp)
CNE4	Intron 7	121730105–121731402	1298	67% (163bp)
CNE5	Intron 1	121563302–121563959	658	75% (221bp)

Ensembl genome browser, to confirm whether these CNEs are overlapping with exons or non-protein coding RNA. It was found that selected CNEs were unique to *GLI2* introns, and do not overlap with any exon or protein coding RNA.

3.2 Identification of co-orthologs in the zebrafish genome

Several studies suggest that after divergence from tetrapods, 450 Mya, zebrafish genome including other teleost members, appears to have undergone duplication (Abbasi *et al.*, 2009, Postlethwait *et al.*, 2000). It is proposed that as a result of this duplication, zebrafish genome has two copies of *gli2* genes, i.e. *gli2a* (chromosome 9) and *gli2b* (chromosome 11) (Ke *et al.*, 2005, Ke *et al.*, 2008). Comparison of the selected subset of human *GLI2*-associated CNEs in zebrafish, using BLAST, it is revealed that the highly conserved element CNE2 (present in the intron 2 of human GLI2 gene, Figure 3.1) has two co-orthologs in zebrafish genome. These duplicated

CNEs (dCNEs) maintained their physical association with *gli2a* and *gli2b* in zebrafish genome, and were assigned to distinct identities, i.e. CNE2a and CNE2b respectively. CNE2a is located in intron-2 of zebrafish *gli2a* gene (Dr-gli2-CNE2a) and CNE2b is present downstream of zebrafish *gli2b* gene (Dr-gli2-CNE2b) (Figure 3.2).

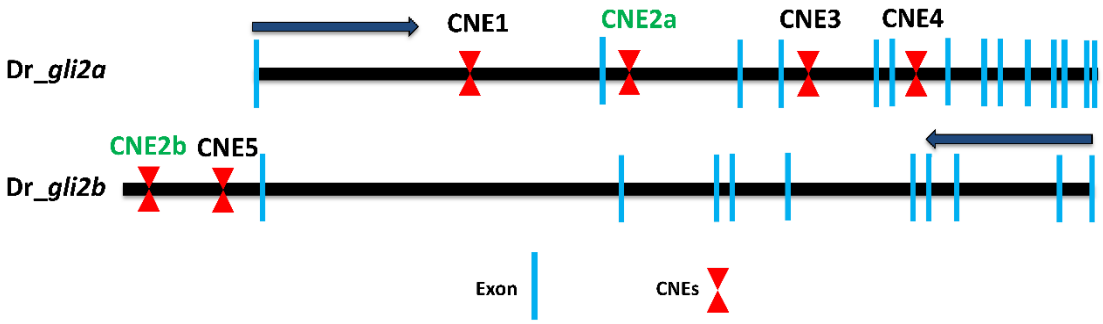

Figure 3. 2 Comparative genic architecture of human *GLI2*-associated CNEs in zebrafish revealed CNE2 is duplicated in zebrafish genome.

Comparative view of zebrafish gli2 genes showing exons and conserved noncoding elements in light blue and red colour respectively. Arrows show the direction of each gene, gli2b is in reverse orientation. CNE2 is duplicated in zebrafish and has two copies. Zebrafish duplicated enhancers are present in the introns of gli2a, and downstream of gli2b. dCNEs (CNE2a and CNE2b) are shown in their respective position and written in green. Dr; Danio rerio

3.3 Association of identified human CNEs to *GLI2* locus by syntenic mapping

To confirm *in silico*, whether these CNEs are truly associated with GLI2 gene, a gene synteny was drawn for *GLI2* and its flanking genes, between tetrapod and teleost orthologous loci (Figure 3.3). On inspection, tetrapod-specific gene syntenic comparison failed to assign precisely the target gene for the selected subset of CNEs. However, comparative syntenic analysis of flanking genes in teleost revealed that synteny breakage occurs for rest of the flanking genes and only *GLI2* gene maintains conserved association with these five CNEs, in fugu and zebrafish (Figure 3.3).

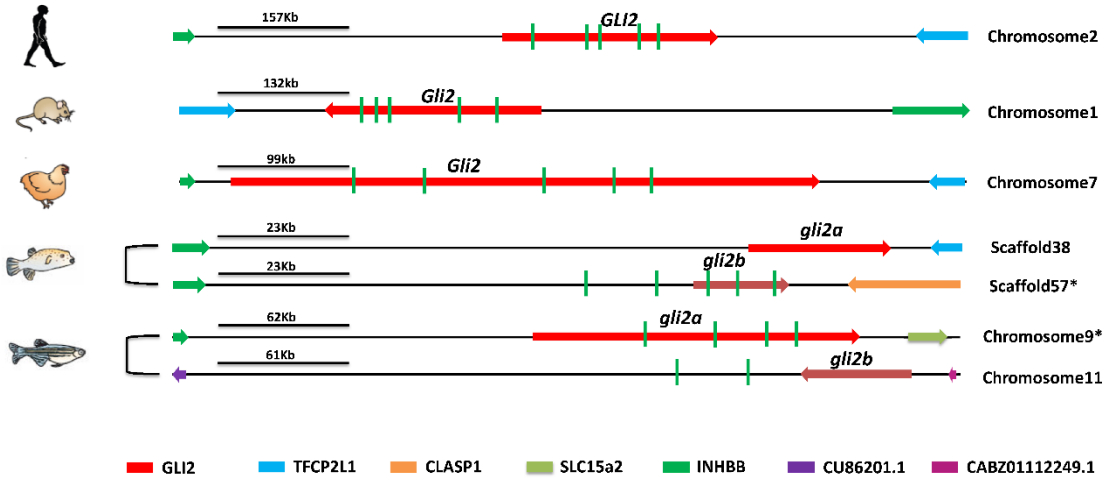

Figure 3. 3 Human *GLI2*-associated CNEs maintain their physical linkage with GLI2 gene

Comparative syntenic analysis of human, mouse, and chicken orthologous loci depicts the conserved presence of three genes GLI2, TFCP2L1, and INHBB in the nearest vicinity of these five CNEs (light green vertical line). Increasing the depth of synteny comparison by including orthologous loci from teleost fish lineage, it became evident that GLI2 maintains its association with the CNEs down to fish. Genes are color-coded. Direction of arrow depicts the direction of gene transcription. Light green vertical line depicts the position of CNE-enhancer. Horizontal black line depicts scale; asterisk (*) mark presents the duplicated copy of gli2 which were used during the comparative analysis.

3.4 *In vivo* functional analysis of *GLI2*-associated CNEs using a co-injection assay in transiently transfected zebrafish embryos

The CNEs that have been identified through comparative genomics were next tested *in vivo* using zebrafish as a model organism. Zebrafish is widely used as an *in vivo* system to elucidate the role of vertebrate regulatory networks as it shares number of development stages with human and other vertebrates, and has transparent embryos which makes it easy to study the reporter gene expression. In order to characterize human *GLI2*-associated CNEs and to decipher their putative activity, a medium throughput approach was first used for *in vivo* functional assay (co-injection), having minimal mouse β-globin promoter and green flourescent protein (GFP) as a reporter gene in zebrafish embryos, as described previously (Woolfe *et al.*, 2005) (Figure 3.4). Two time points were selected to study the putative activity of

these elements, i.e. 24 and 48 hours post fertilization (hpf), as at these stages the zebrafish embryos are most visible and completely formed (Kimmel *et al.*, 1995). To note this study is done to test independent enhancing function of CNEs in zebrafish autonomous of their genomic context. Each CNE was amplified from genomic DNA of human and zebrafish, followed by its purification. Purified amplicons of each CNE alongwith amplified reporter gene cassette were then injected together at 1-2 cell stage of developing zebrafish embryos (see Materials and Methods). The reporter gene casette containing the minimal β-globin promoter and GFP was amplified from β-globin EGFP vector (Figure 3.4). Co-injection approach exploits the transparency and fast growth of zebrafish embryos and has shown its potential for functionally testing enhancer elements among conserved non-coding regions (Abbasi *et al.*, 2007, Woolfe *et al.*, 2005). Reporter gene activity by all the five CNEs in multiple independent transient transgenic assays showed reproducible tissue-specific expression after 24 hpf (day-2) and 48 hpf (day-3). Reporter gene cassette with non-conserved randomly selected region from genomic region were tested as a negative control and did not evoke any sort of reporter gene activity at approximately 24 as well as 48 hpf.

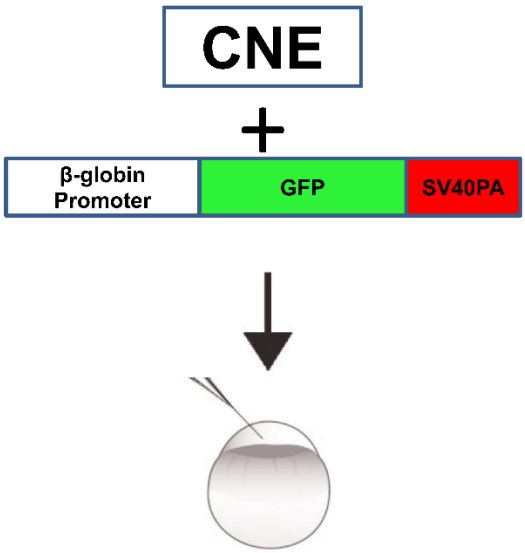

Figure 3. 4 Schematic representation of co-injection strategy used for *in vivo* characterization of CNEs

Each CNE and reporter cassette containing β-globin, GFP and SV40 poly A tail were amplified using standard PCR protocol (see details in Material and Methods) and injected with phenol red into zebrafish embryo at 1-2 cell stage.

3.4.1 CNE1

CNE1 (933 bp) is located in intron 1 of GLI2 gene and directs GFP expression largely in various subdivisions of CNS, i.e. forebrain (8% of expressing embryos(EE), midbrain (3% of EE), hindbrain (19% of EE) (Figure 3.5A and B), and spinal cord (6% of EE) on day-2 of development (~24 hpf). In addition to CNS, GFP expressing domains for CNE1 on day-2 of development are muscle (17% of EE) (Figure 3.5C), blood (13% of EE), and epidermis (6% of EE) (Figure 3.5D). Blood precursor cells show reporter gene expression at day-2 of development and later on their expression is reduced by day-3 (Figure 3.5F). Cardiac and circulating blood cells also were observed to show GFP expression, 3% and 13 of EE % respectively (Figure not shown due to regions being

difficult to image). CNE1 induces most remarkable GFP reporter gene expression in the developing hindbrain (45% expressing embryos, Figure 3.5E) at day-3 of development (~48 hpf), as compared to day-2 (19% of expressing embryos). In addition to CNS and hindbrain, reporter gene expression is also observed in epidermis/skin (15% of expressing embryos).

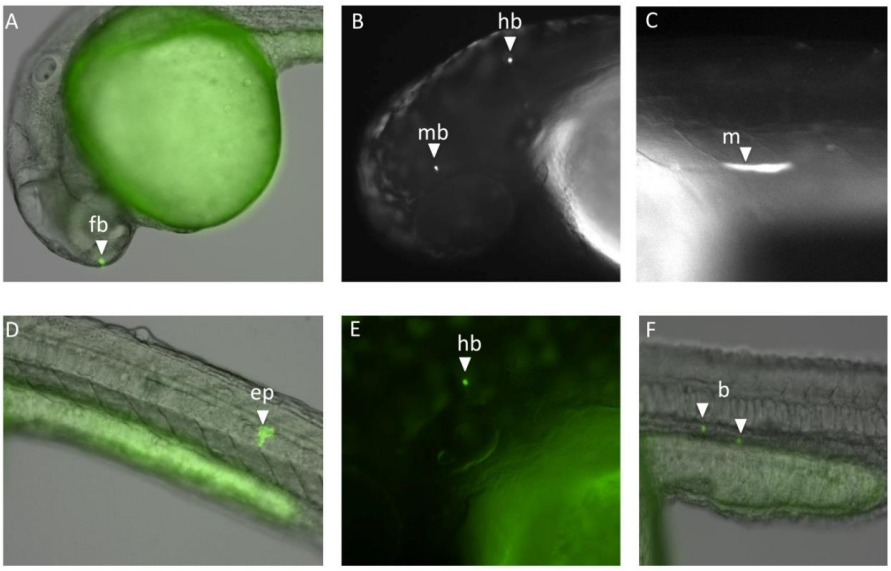

Figure 3. 5 CNE1 mediated reporter gene expression prediminantly in brain and ventral caudal region

GFP expression is shown in live embryos at ~24 hpf and ~48 hpf, compound (combining bright field and fluorescence views) (A, D, E, and F), and GFP fluorescence is shown in live embryos (B and C). Presented here are GFP expressing cells, indicated by arrowheads, in (A) forebrain in live embryo, (B) GFP expression shown in midbrain and hindbrain, (C) GFP expression in muscles, (D) epidermal cells in the trunk region, (E) GFP expressing primary neurons in hindbrain, (F) Expression seen in unidentified region between yolk sac and vertebral column (blood precursor cells). fb, forebrain; mb, midbrain; m, muscle; hb, hindbrain; ep, epidermis; b, blood precursor cells.

3.4.2 CNE2

Highly conserved CNE2 is positioned in the intron 2 of GLI2 gene, has 531 bp of length, and directs reporter gene expression exclusively in notochord at ~24 hpf

(Figure 3.6A and B) as well as ~48 hpf (Figure 3.6C and D). Reporter gene expression was not observed in any other embryonic domain at both the screening time points (24 and 48 hpf).

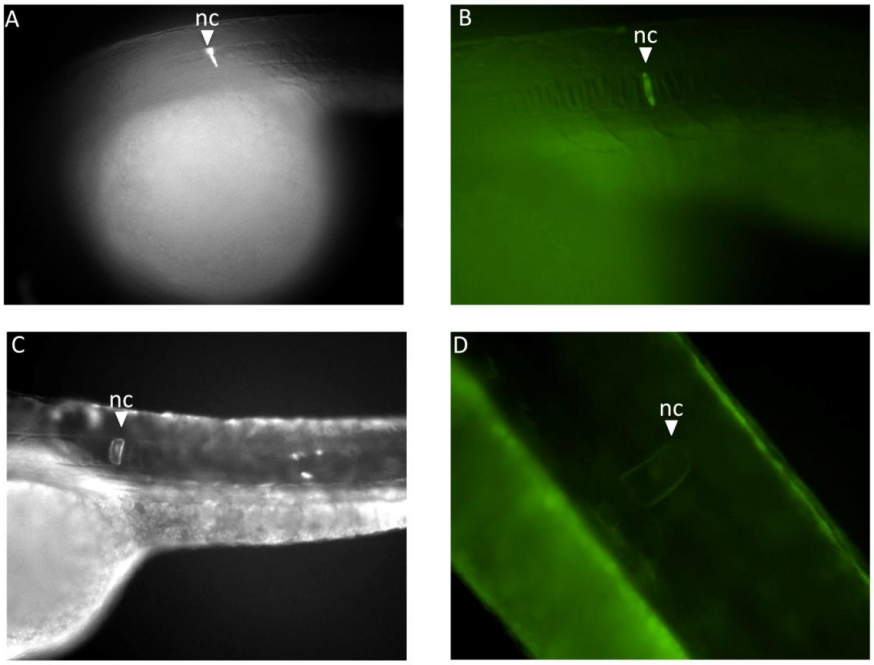

Figure 3. 6 CNE2 exclusively regulates GFP expression in notochord at day-2 and at day-3

GFP expression is shown in live embryos, Fluorescent (A and C), and GFP expression is shown in live embryos (combining bright field and fluorescence views) (B and D). Presented here are GFP expressing cells indicated by arrowheads in notochord (A and B) at day-2 (~24 hpf), (C and D) at day-3 (~48 hpf). nc, notochord.

3.4.3 CNE3

CNE3 (850 bp) is present in intron 4 of *GLI2* gene and directs GFP expression mainly in the epithelium of otic vesicle (48% of EE) (Figure 3.7A), muscle fibers (52% of EE) (Figure 3.7B) and hindbrain (11% of EE), at ~24 hpf. CNE3 also evokes reporter gene expression in blood precursor cells (Figure 3.7B). Moreover, GFP expression is very high in the branchial arch (46% of EE) at day-2 (Figure 3.7C). At ~48 hpf, the main GFP expression domains for CNE3 are hindbrain (65% of EE) (Figure 3.7D), otic

vesicle (25% of EE), and muscle fibres (25% of EE). The expression of CNE3 in the branchial arch is also significant at day-3 (40% of EE), along with cardiac cell showing reporter gene expression (5% of EE).

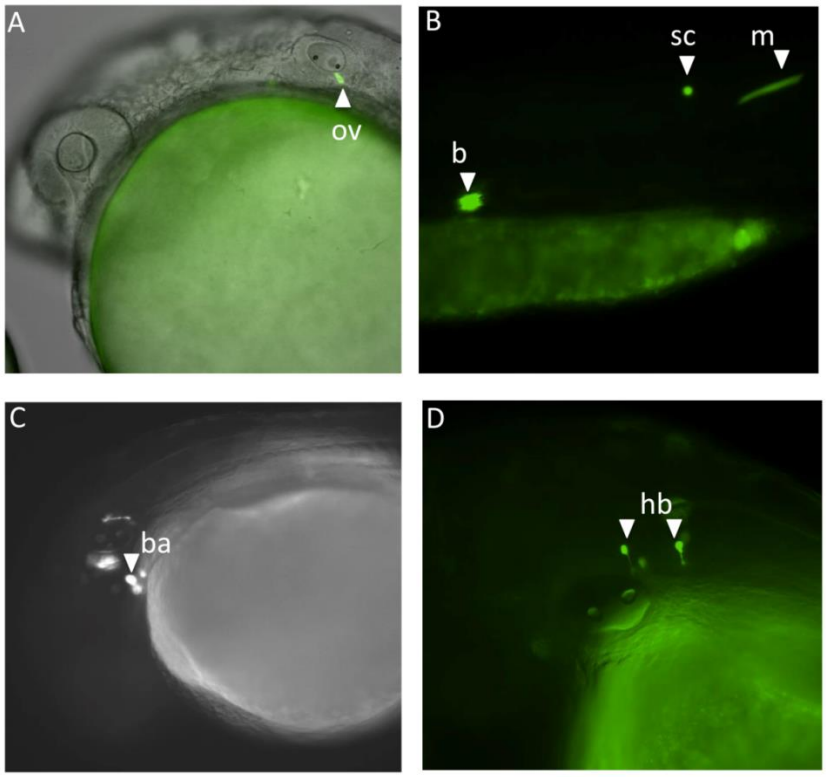

Figure 3. 7 CNE3 mediated reporter gene is expressed predominantly in otic vesicle, hindbrain and muscle cells.

GFP expression is shown in live embryos, compound (combining bright field and fluorescence views) (A, B, and D), and fluorescent embryos (C). Arrowheads indicate GFP expressing cells. Presented here are (A) GFP expressing cells in otic epithelium (~24 hpf), (B) GFP expression shown in spinal cord, muscle cells and unidentified domain between vertebral column and yolk sac, (C) Branchial arch, (D) GFP expressing neurons in hindbrain at 48 hpf. ov, otic vesicle; sc, spinal cord; m, muscle; hb, hindbrain; b; blood precursor cells; ba, branchial arch.

3.4.4 CNE4

CNE4 (1298 bp) is located in intron 7 of *GLI2* gene and induces reporter gene expression explicitly in muscle cells at day-2 (93% of EE) and day-3 (96% of EE). None

of the other domains show reporter gene expression in developing embryos at ~24 hpf, as well as at ~48 hpf (Figure 3.8A and B).

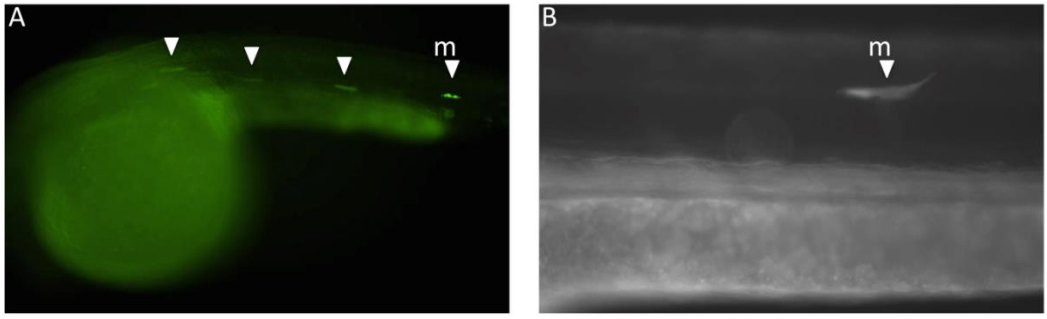

Figure 3. 8 CNE4 specifically regulates GFP expression in muscle cells at day-2 and at day-3

GFP expression is shown in live embryos, compound picture (combining bright field and fluorescence views) at ~24 hpf (A), fluorescent (B) at ~48 hpf. Arrowheads indicate GFP expressing cells. m; muscle fibres.

3.4.5 CNE5

CNE5 resides at the start of intron 1 of *GLI2* gene and is of 658 bp length. CNE5 drives GFP reporter gene expression mainly in the hindbrain (23% of EE) (Figure 3.9A), spinal cord (28% of EE) and muscle cells (34% of EE) (Figure 3.9B), at ~24 hpf. Reporter gene expression is also detected in branchial arch (37% of EE) and blood precursor cells. Approximately 3% of the embryos demonstrate GFP expression in the circulating blood and notochord. The highest reporter gene expression at day-3 is detected in spinal cord (53% of EE) (Figure 3.9D). In addition to spinal cord, GFP expression is also detected in hindbrain (Figure 3.9C) and muscle cells, i.e. 40% and 30% of EE respectively.

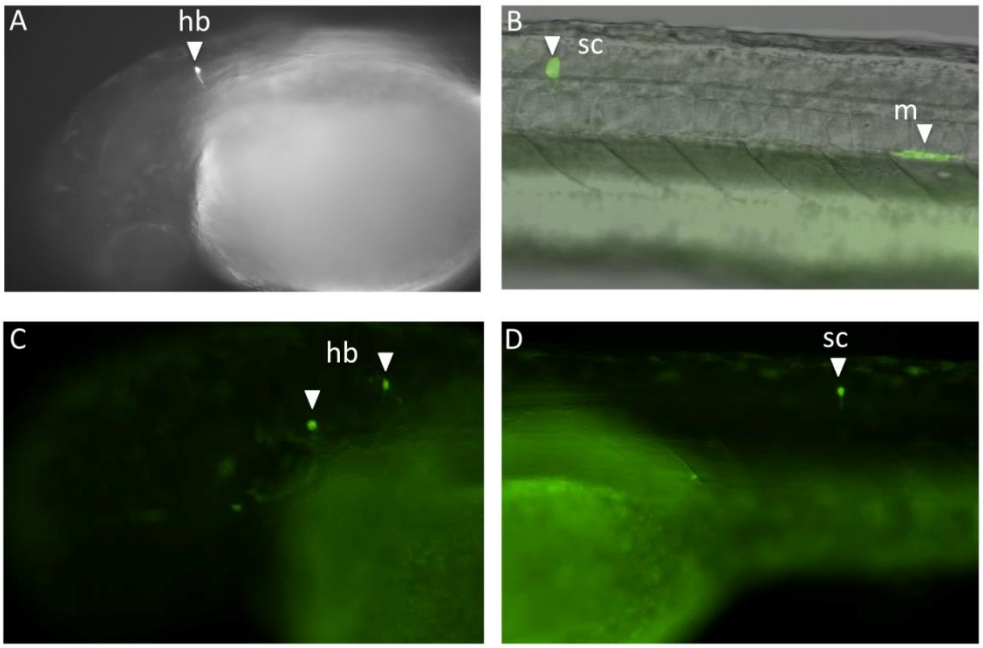

Figure 3. 9 CNE5 drives GFP expression in CNS.

GFP expression is shown in live embryos, fluorescent (A) and compound view (combining bright field and fluorescence views) (B, C and D). Arrowheads indicate GFP expressing cells. Presented here are (A) GFP expressing cells in hindbrain (~24 hpf); (B) GFP expression shown in spinal cord and muscle (~24 hpf); (C) GFP expression in hindbrain (~48 hpf); (D) neuron in spinal cord showing GFP expression (~48 hpf). hb, hindbrain; sc, spinal cord; m, muscle.

3.4.6 Zebrafish duplicated CNEs (CNE2a and CNE2b)

As found by VISTA analysis, highly conserved *GLI2*-associated CNE2 is present in intron 2 of human *GLI2* gene and has two co-orthologs in zebrafish genome due to lineage specific duplication. These co-orthologs are retained after duplication at *gli2a*/*glib* gene in zebrafish genome. These co-orthologs share high sequence similarity within and between both lineages (human and zebrafish) (Figure 3.10A). To investigate the functional aspects of zebrafish duplicated CNEs, Dr-*gli2*_CNE2a (207 bp) and Dr-*gli2*_CNE2b (208 bp) were amplified from zebrafish genomic DNA and injected into zebrafish embryos using co-injection strategy. Dr-*gli2*_CNE2a and Dr-

*gli2*_CNE2b showed similar GFP expression at 48 hpf in the primary neurons of the hindbrain, 2.3% and 3.6% of EE respectively (Figure 3.10 B and C).

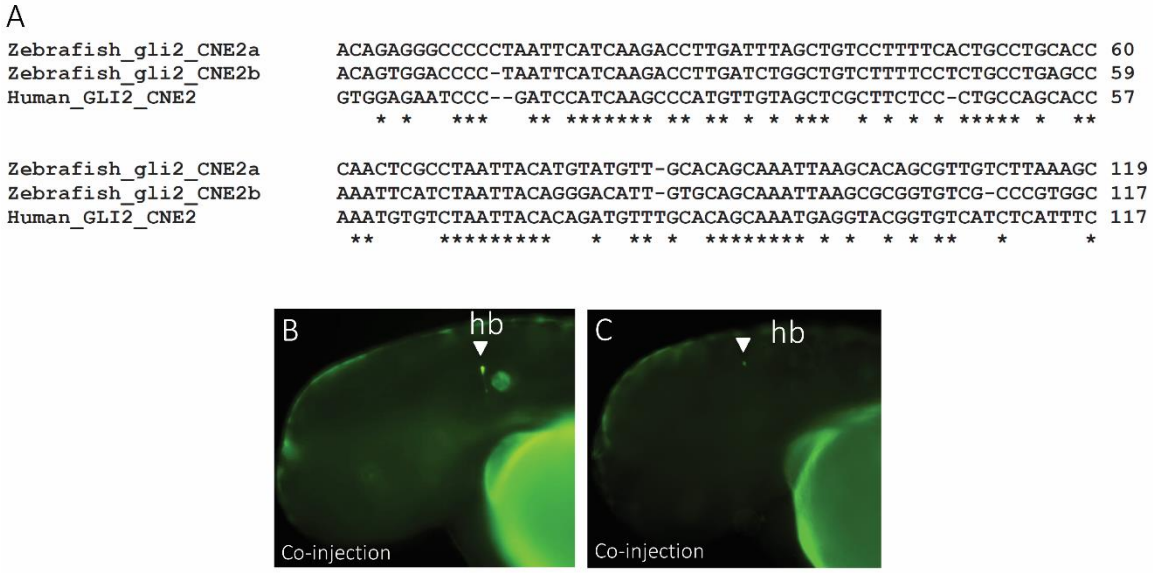

Figure 3. 10 Sites of GFP expression induced by duplicated orthologous sequences of CNE2 in zebrafish embryos.

(A) Multiple sequence alignment of GLI2-associated CNE2 and its co-orthologous sequences from zebrafish CNE2a and CNE2b. (B) Dr-gli2_CNE2a directed GFP expression in hindbrain (C) Dr-gli2_CNE2b drives GFP expression in the hindbrain region. hb, hindbrain.

3.4.7 Comprehensive outline of GFP expression territories in zebrafish embryos at 24 hpf or 48 hpf

Due to mosaic nature of co-injection assay, approximately 200 zebrafish embryos were microinjected and screened at day-2 and day-3 for each element, to derive an overall spatio-temporal representation of enhancer activity through cumulative data from all GFP expressing embryos. Combined graphical expression data for each CNE was compressed into a JPEG file together with graphical description of expression domains. Percentages and categories of cell types that are positive for each CNE are highlighted as a schematic representation shown in Figure 3.11.

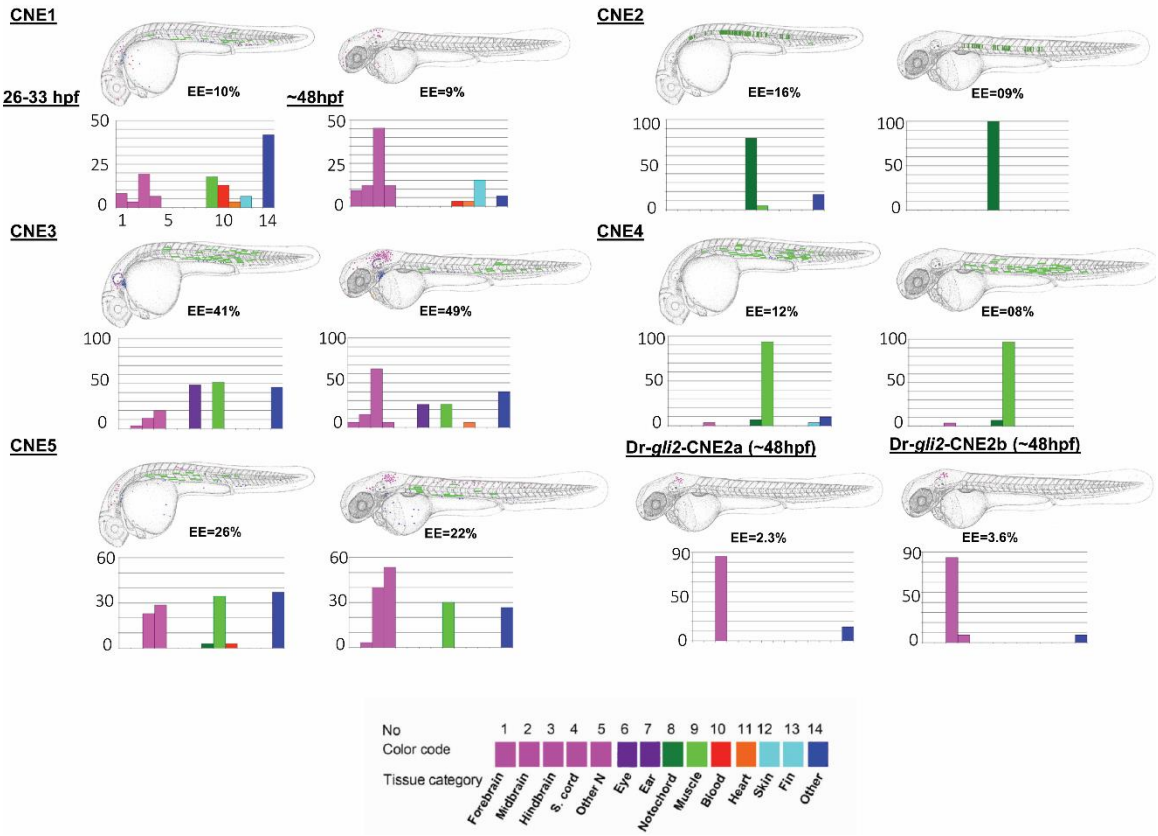

Figure 3. 11 Sites of GFP expression induced by *GLI2*-associated CNEs in zebrafish embryos

Sites of GFP signals recorded in zebrafish embryos transiently transfected with a construct, in which the reporter gene was induced by individual GLI2-associated CNEs (indicated by name), are depicted in schematic representations of day-2 (24 hpf) or day-3 (48 hpf) embryos. The percentage of positive embryos per CNE (EE) are indicated for each of the constructs. Categories of cell type that were positive for a given element are colour coded, with each dot representing a single GFP-expressing cell. Bar graphs display the percentage of GFP-expressing embryos that show expression in each tissue category for a given element. Bar graphs use the same colour code as the schematics for each cell type.

3.5 Study of *GLI2*-associated CNEs by *Tol2* based transgenic assay

To verify the results from co-injection assay, and to examine the potential *cis*-regulatory function of *GLI2*-associated CNEs in detail, we used a more robust and efficient strategy based on *Tol2* vector with *c-fos* minimal promoter (Figure 3.12). *Tol2* based transgenesis has the advantage of stronger reporter expression with

reduced mosaicism due to efficient and stable integration in the zebrafish genome as compared to co-injection assay (Kawakami *et al.*, 2004, Fisher *et al.*, 2006).

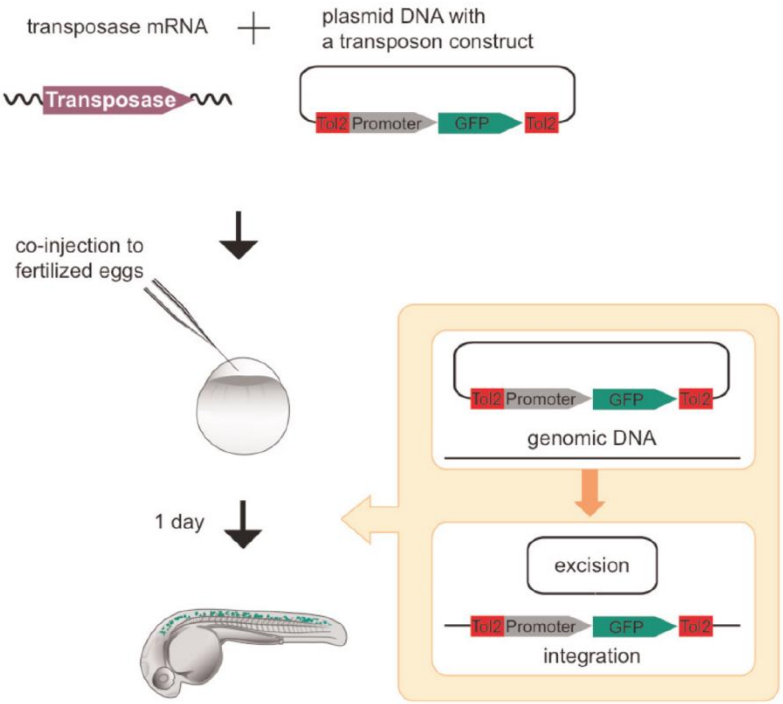

Figure 3. 12 Transient expression assay in zebrafish.

The transposase mRNA and a transposon donor plasmid containing a Tol2 construct with the putative element and the gene encoding green fluorescent protein (GFP) are co-injected into zebrafish fertilized embryos. The Tol2 construct is excised from the donor plasmid and integrated into the genome of somatic cells. The activity of the enhancer can be visualized in the injected embryo, because GFP is expressed predominantly in a region where the enhancer/promoter should activate transcription. (Adapted from Kawakami 2007).

The injected embryos were screened (~24 and ~48 hpf) for comparable GFP expression with the co-injection assays. Due to position effects, random and multiple integration of *Tol2* in the zebrafish genome, the pattern of GFP reporter expression of a given construct may differ frequently. So the expression patterns with more than 25% of expressing embryos were considered as positive for each of the construct. Moreover, zebrafish embryos that showed significant and consistent GFP

expression in other domains which were not observed in co-injection assay were also included in this data.

3.5.1 CNE1 evokes GFP expression in epidermis with the *Tol2* reporter system

CNE1 drives GFP expression in the CNS (especially hindbrain) and tail bud region (blood precursor cells) with co-injection strategy. Using *Tol2* transgenic assay, CNE1 does not reproduce significant reporter gene expression in the hindbrain or sub-domains of CNS, whereas very few cells were detected with GFP expression in the ventral side of the tail bud region. However, GFP expression in the skin cells by co-injection (8% of EE) (Figure 3.13A) is recapitulated by the *Tol2* transgenic system (21% of EE) at ~48 hpf (Figure 3.12B), which correlates with the endogenous expression of *GLI2* gene.

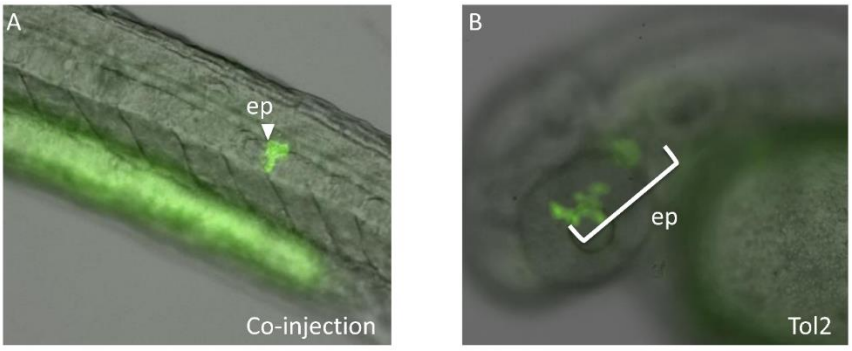

Figure 3. 13 Comparison of GFP expression in epidermis by co-injection and *Tol2* transgenic system.

GFP reporter gene expression of CNE1 by (A) co-injection assay using β-globin promoter and (B) Tol2 transgenic system using c-fos promoter. Images of live zebrafish embryos at (A) 24 hpf and (B) 48 hpf. Arrowhead and marked area indicate GFP expressing cells. Both assays showed expression in skin cells. ep, Epidermis.

3.5.2 CNE2 induces reporter gene expression in CNS and pectoral fin

CNE2 activates reporter gene expression categorically in notochord with co-injection transgenesis. Similar reporter gene expression was recaptured at day-2 and day-3 when *Tol2* transgenic system is used (Figure 3.14C). In addition to notochord, some other domains also indicate consistent and reproducible GFP expression of CNE2 with *Tol2* transgenesis. Most notable activity domains for CNE2 are spinal cord (73% of EE), and muscle cells (Figure 3.14A). Moreover, CNE2 induces dominant reporter gene expression in epidermis (Figure 3.14B).

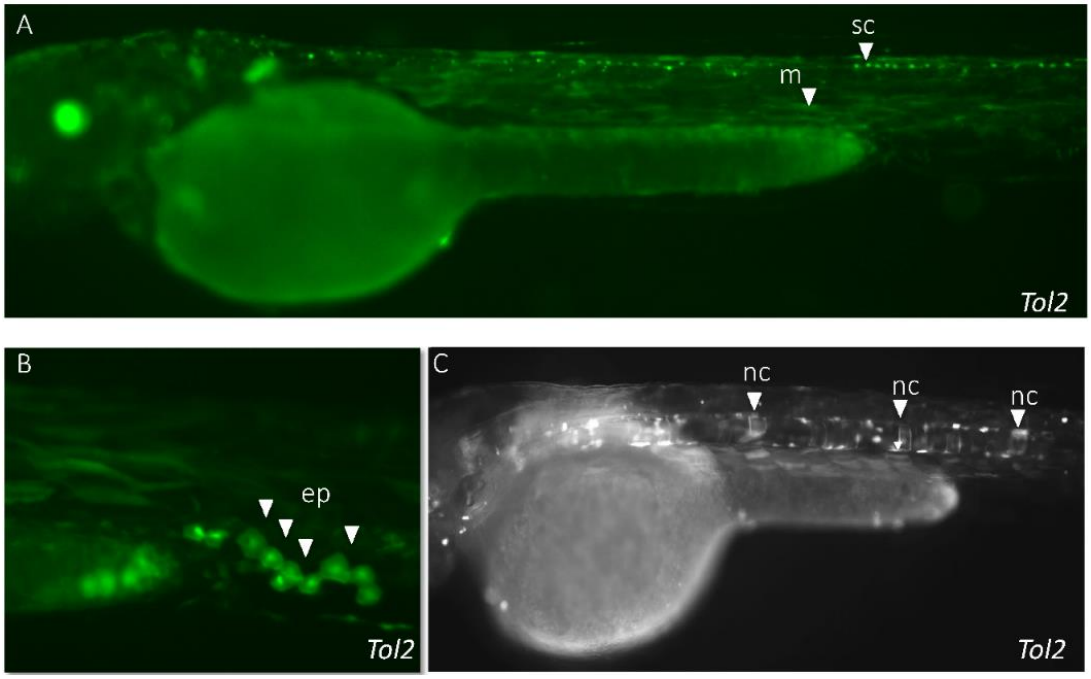

Figure 3. 14 CNE2 drives GFP expression mainly in the spinal cord, epidermis and notochord by *Tol2* transgenic system.

Reporter gene expression by human GLI2 CNE2. Images of live zebrafish embryos 48 hpf (A-B) and 24 hpf (C), lateral views, anterior to left, dorsal to top. Arrowheads and marked area indicate GFP expressing cells. (A) GFP expression in the spinal cord and muscle cells. (B) GFP expression in epidermis. (C) GFP expression shown in the notochord. nc, notochord; sc, spinal cord; m, muscle; ep, epidermis.

CNE2 also evokes consistent and reproducible reporter gene expression in the neurons of hindbrain (60% of EE) (Figure 3.15A). In order to test the hypothesis whether the expression of CNE2 (hindbrain enhancer) is rhombomere-specific of developing zebrafish embryos, it was tested in a transgenic zebrafish *krox20* line. *Krox20* is a gene which is expressed in r3 and 5, and this transgenic line is frequently used to study rhombomere-specific expression of transcription factors (Parker *et al.*, 2011). This transgenic line shows RFP expression as a marker in rhombomere 3 and rhombomere 5 (Distel *et al.*, 2009). CNE2 mediated GFP expression is detected in and around r3 and r5, as well as in the entire hindbrain territory indicating that this enhancer does not regulate GFP expression in certain rhombomeres (Figure 3.15B).

In addition, with *Tol2* reporter system, CNE2 also showed robust and reproducible GFP expression in the developing pectoral fin (36% of EE) at ~48 hpf (Figure 3.15C and D). Reporter gene expression in fin is detected even at day-4, i.e. 72 hpf (Figure 3.15 C). The core-conserved region of CNE2 for tetrapod/teleost was also injected to study the effect of minimal region of this element. Therefore, the core-conserved region (208 bp) was cloned into *Tol2* vector. This truncated CNE2 was shown to drive GFP expression almost similar to full length CNE2 in notochord (Figure 3.16A) and hindbrain (Figure 3.16B). However, reporter gene expression was not observed in the pectoral fin at any developmental stage.

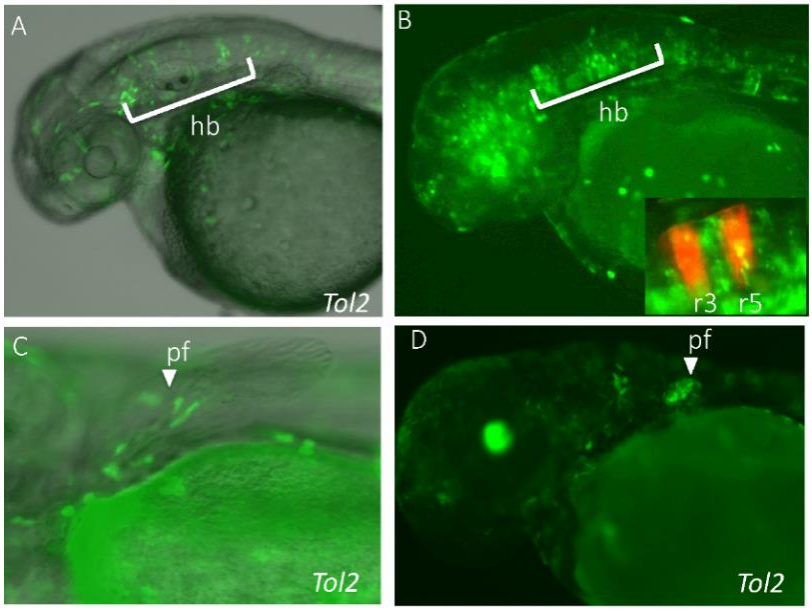

Figure 3. 15 CNE2 drives GFP expression in the hindbrain and fin by *Tol2* reporter system.

Images of live zebrafish embryos (A, B, and D) at ~48 hpf, lateral views, and anterior to left and (C) at 72 hpf, ventral view. Marked areas indicate GFP expressing cells. (A) GFP expression is in the primary neurons of the hindbrain. (B) The expression territory is confirmed in the entire hindbrain as determined by comparison in a krox20 transgenic line showing r3r5-RFP expression (inset). (C and D) CNE2 induces GFP expression in the developing pectoral fin. hb, hindbrain; r3, rhomobomere3; r5, rhomobomere5; pf, pectoral fin.

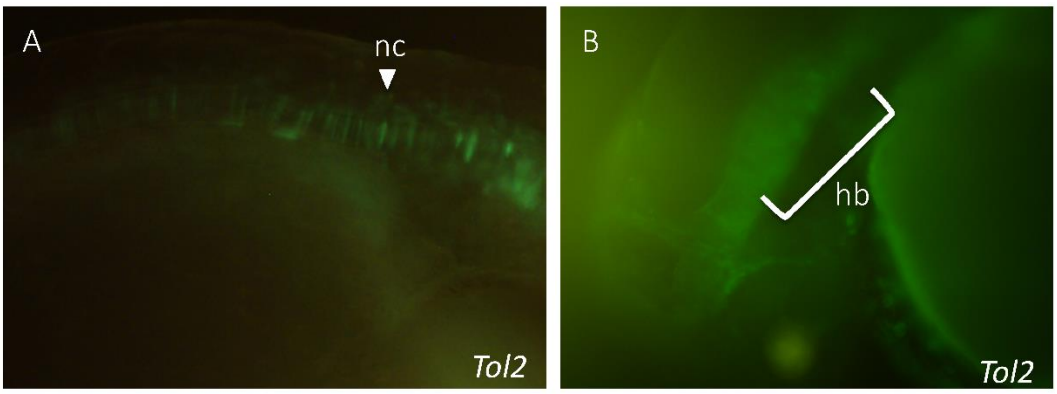

Figure 3. 16 Truncated *GLI2* CNE2 (208bp) triggers GFP expression in notochord and hindbrain.

Images of live zebrafish embryos 24 hpf (A) and at 48 hpf (B), lateral views, anterior to left, dorsal to top. Arrowhead and marked area indicates GFP expressing cells. Truncated CNE2 induces GFP expression in the (A) notochord and (B) hindbrain.nc; notochord, hb; hindbrain.

3.5.3 CNE3 governs GFP expression in CNS by *Tol2*

As discussed above in zebrafish co-injection assay, CNE3 triggers GFP signal remarkably in CNS. The most prominent reporter gene expression is observed in neurons of hindbrain and spinal cord. CNE3 reproduces a prominent and similar pattern of GFP expression at 24 and 48 hpf in the neuronal cells of CNS in zebrafish embryos with *Tol2* based transgenic system. Most prominent GFP expression is induced by CNE3 in the hindbrain (63% of EE) (Figure 3.17A and B), spinal cord (41% of EE) (Figure 3.17E), otic vesicle (36% of EE) (Figure 3.17F), and muscle cells (82% of EE) (Figure 3.17E) at 48 hpf. Similar to CNE2, *Tol2*-CNE3 construct was then injected in transgenic zebrafish *krox20* line, which drives RFP expression in rhombomere 3 and 5, to check for rhombomere-specificity. CNE3 was found to drive GFP expression in and around r3 and r5 (inset of Figure 3.17B). Moreover, CNE3 also evokes reporter gene expression in the trigeminal cranial ganglia of zebrafish embryos (Figure 3.17C and D), while GFP expression by CNE3 in the heart and circulating blood is as well.

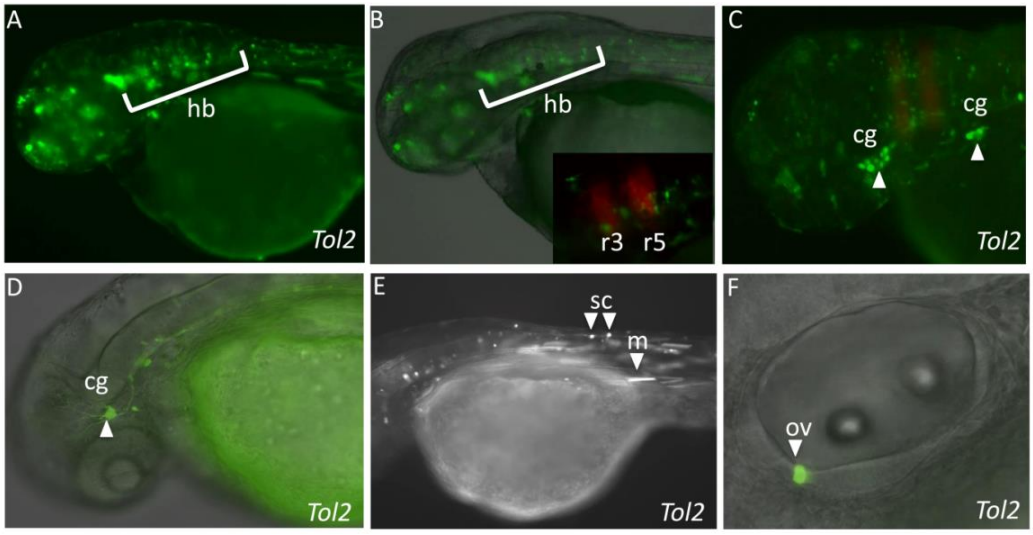

Figure 3. 17 CNE3 controls GFP expression in the central nervous system and sensory organs of zebrafish embryos.

Images of live zebrafish embryos at 48 hpf, lateral views, anterior to left, dorsal to top. Arrowheads and marked area point to GFP expressing cells. (A and B) CNE3 induced GFP expression in hindbrain. GFP expression in the hindbrain by CNE3 in transgenic line krox20 shows RFP expression in and around rhombomere 3 and 5 (inset B). (C and D) Expression in the cranial ganglia. (E) CNE3 drives GFP expression in the spinal cord and muscle at 48 hpf. (F) CNE3 is expressed in the otic vesicle. Hb, hindbrain; cg, cranial ganglia; sc, spinal cord; m, muscle; ov, otic vesicle; r3, rhomobomere3; r5, rhomobomere5.

3.5.4 CNE4 induces GFP expression limited to muscle cells

CNE4 drives GFP expression exclusively in the muscle cells with co-injection assay. However, with *Tol2* reporter system very few muscle cells were detected with GFP expression (Figure 3.18A), which are less than the number of GFP-positive muscle cells as compared to other CNEs (CNE2 and CNE3). However, few skin cells were also detected with GFP signals (Figure 3.18B). Regulatory activity variation by two different strategies suggests that such elements studied further using different minimal promoters such as E1b.

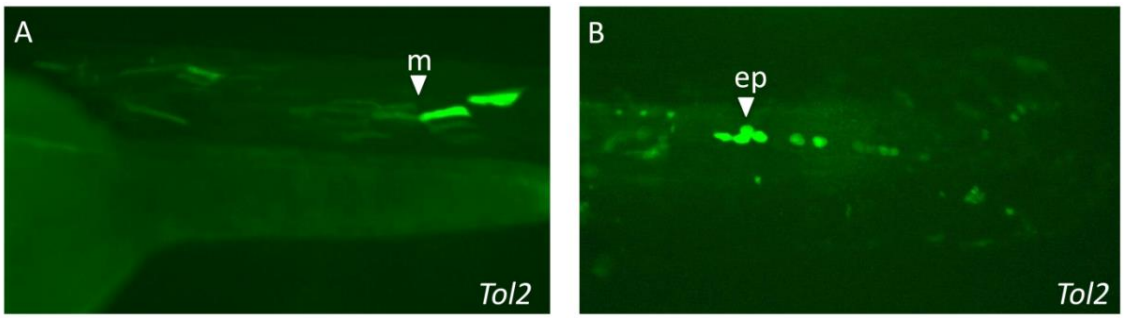

Figure 3. 18 CNE4 induces non-significant GFP expression by *Tol2* transgenic system.

Images of live zebrafish embryos 48 hpf (A-B), lateral views, anterior to left, dorsal to top. Arrowheads indicate GFP expressing cells. CNE4 induced GFP expression in (A) muscle cells and (B) epidermis. m, muscle; ep, epidermis.

3.5.5 CNE5 induced GFP expression in CNS and pectoral fin by *Tol2* transgenesis

CNE5 acts as a very strong enhancer element to induce GFP expression in the CNS with co-injection assay. Using *Tol2* assay, CNE5 triggers a robust pattern of GFP expression in two major domains of CNS, i.e. hindbrain (58% of EE) (Figure 3.19A) and spinal cord (65% of EE) (Figure 3.19E), at ~48 hpf. Similar to CNE2 and CNE3, CNE5 was tested in *krox20* line showing GFP expression in and around r3 and r5 (inset of Figure 3.19B). Other than CNS, the prominent domains on day two of development for CNE5 are muscle (48% of EE) (Figure 3.19E), blood (27% of EE), and mesoderm (23% of EE) (Figure 3.19D). Another prominent domain for CNE5 is the developing pectoral fin (31% of EE) (Figure 3.19C and D). Reporter gene activity was also detected in the heart (15% of EE) of the zebrafish embryos.

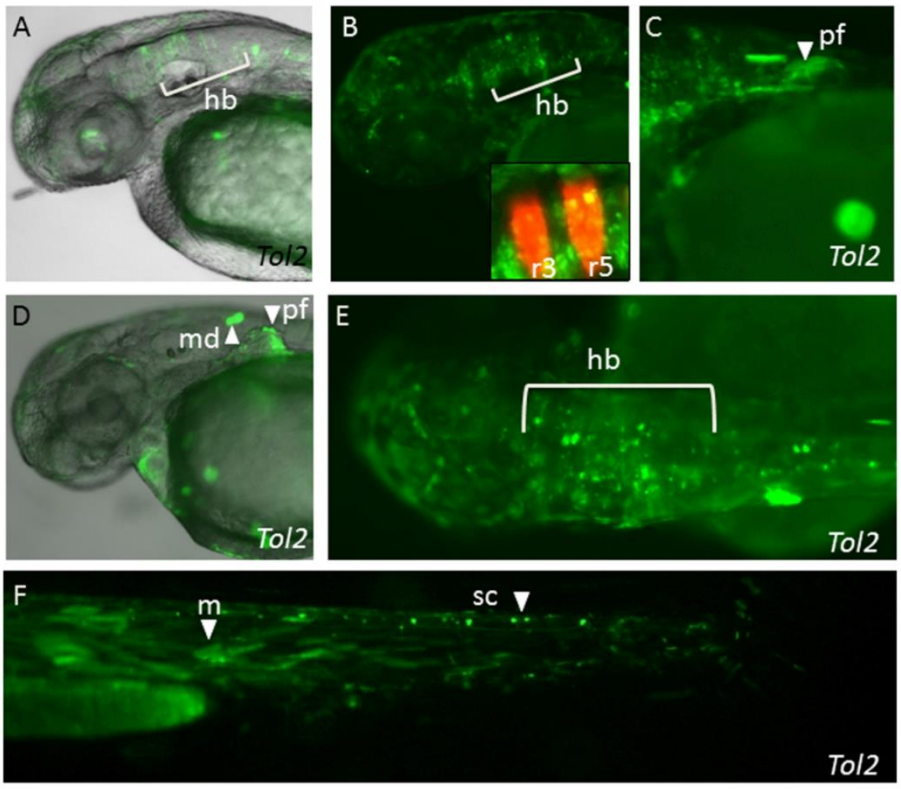

Figure 3. 19 CNE5 mediated reporter gene expression was more prominent in hindbrain and spinal cord.

GFP expression is shown in live embryos at 48hpf. Arrowheads and marked area indicate GFP expressing cells. (A) GFP expression is shown in hindbrain. (B) GFP expression in the hindbrain is compared by transgenic krox20 line, which shows RFP expression in and around rhombomere 3 and 5 (inset B). (C) GFP expression in pectoral fin. (D) Expression in the pectoral fin and mesoderm in the dorsal head region. (E) GFP expressing neuronal cells in the dorsal hindbrain. (F) GFP expressing neurons in spinal cord and muscle cells. hb, hindbrain; md, mesoderm; m, muscle.

3.5.6 Duplicated CNEs (CNE2a and CNE2b) have similar GFP expression in hindbrain and pectoral fin with *Tol2* transgenesis

Zebrafish co-orthologs (Dr-*gli2*_CNE2a and Dr-*gli2*_CNE2b) of human *GLI2*-CNE2 have a similar expression profile, as shown above, in the hindbrain territory by co-injection assay. Dr-*gli2*_CNE2a and Dr-*gli2*_CNE2b were then injected further in zebrafish embryos using *Tol2* based transgenic assay. These dCNEs apparently also exhibited similar GFP expression predominantly in the hindbrain by *Tol2* transgenic system. CNE2a triggers GFP expression in hindbrain (33% of EE) (Figure 3.20A and B) and spinal cord (48% of EE) (Figure 3.20E), whereas CNE2b drives GFP expression in the hindbrain (25% of EE) (Figure 3.20F) and muscle cells (60% of EE) (Figure 3.20G). In addition, dCNEs revealed a robust and reproducible reporter expression in the developing pectoral fin of zebrafish embryos (Figure 3.20C and H). As compared to CNE2a, CNE2b was unable to induce GFP signal significantly in the spinal cord. So as compared to rest of the CNEs (2, 3, 5 and 2a), it was considered negative for spinal cord. Duplicated CNEs have similar GFP expression pattern in the muscle cells as well (Figure 3.20D and G).

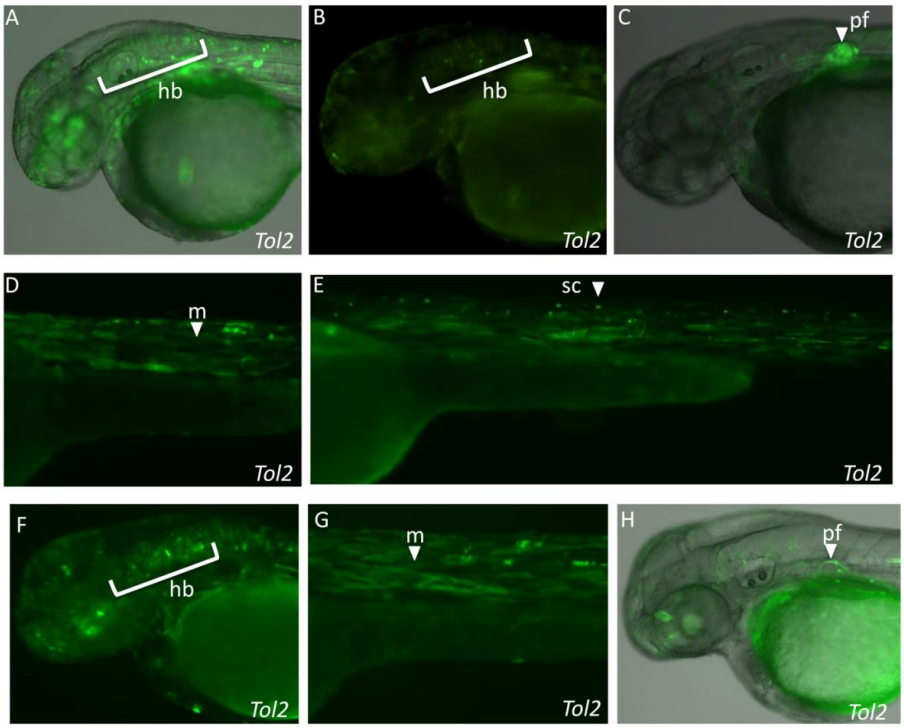

Figure 3. 20 Dr-*gli2*_CNE2a and Dr-*gli2*_CNE2b mediated reporter gene expression remained more prominent in hindbrain and pectoral fin

GFP expression is shown in live embryos at 48hpf. Arrowheads and marked area indicate GFP expressing cells. (A-E) GFP expression is shown by Dr-gli2_CNE2a. (A-B) GFP expression induced in hindbrain, (C) GFP expression in pectoral fin, (D) GFP expression in the muscle cells, (E) GFP expression in primary neurons of spinal cord. (F-H) GFP expression shown by Dr-gli2_CNE2b. (F) GFP expression in the hindbrain, (G) GFP expression in the muscle cells, (H) GFP expressing cells in pectoral fin. hb, hindbrain; m, muscle; f, fin; sc, spinal cord.

3.6 *In silico* mapping of conserved transcription factor binding sites (TFBSs) within each CNE

Binding motifs for TFs are generally short and degenerate and thus can be encountered frequently in the genome. Individual TFBSs are more conserved than surrounding DNA sequences. TFBSs motif searches have been combined with phylogenetic footprinting of CNEs across different species, to search within CNEs for conserved putative TFBS modules (Loots & Ovcharenko, 2004). TFBSs were identified by MEME tool using the parameters discussed in the Materials and Methods section. Human to teleost pairwise sequence comparison was used to find the conserved

motifs at such an extreme phylogenetic distance. For example, CNE1 that resides within intron 1, and has 993bp length, highly conserved sequence track of ~129bp, almost 85% identical in human/rodents and with ~76% sequence identity in human/fugu (Figure 3.21).

```
Human           TCATCAGTGGGGC--CTTAAA-GATAATACACTTTTTGTTTCAGGCAATTACTTAATCTT 57
Mouse           -------CAGGCC--CTTAAAAGATAGTACAGTTTTTGTTTAAGGCAATTACTTAATCTT 51
Fugu            ---TCACCAGCCCGTCTGAGCAGATAATAAACTTCTGCTTTCAGGTAATTACTTAATCTG 57
                         *  *  ** *   **** ** * ** *  *** *** *************

Human           GACAAACTCCTGTGTTAAGCAGGGCTATTAG---CGGGGTCCATCATTAATCAG-CGCAG 113
Mouse           GACAAACTCCTGTGTTAAGCAGGGCTATTAA---TGGGGCCCACCATTAATCAG-TGCGA 107
Fugu            GACAAACCCGCGTGTTAAGCAGTGCTATTAAGCGCGGGG---GCGATAAATCAGGCACGC 114
                ******* *  *********** *******     ****      ** ******   *

Human           AACC-AAGTGGGTGCTT----- 129
Mouse           ACCC-AGGTGGGTGCTTTCTCT 128
Fugu            ATCGGAGGTGGGTGATT----- 131
                * *  * ******* **
```

Figure 3. 21 Multiple alignments show a human/fish core conserved sequence track within CNE1

CNE1 contains highly conserved sequence track of 129bp between human/fugu. Spaces in between the alignment indicate deletion events, while a star symbol underneath represents a nucleotide position conserved in all lineages.

The TFBSs identified by MEME tool are color coded, and have binding sites for multiple developmentally important transcription factors (Figure 3.22). Publicly available online tool STAMP (Mahony & Benos, 2007) was used to study these motifs further against the TRANSFAC v11.3 library (Figure 3.22) (details are given in Material and Methods). Many of these putatively identified transcription factors (Table 3.2) were found to be key developmental regulators, and known to be co-expressed with Gli2 during early embryonic development as verified from the Mouse Genome Informatics database (http://www.informatics.jax.org/). The binding site for each transcription factor was also confirmed manually by multiple alignment of human-teleost sequences with CLUSTALW (Figure 3.23).

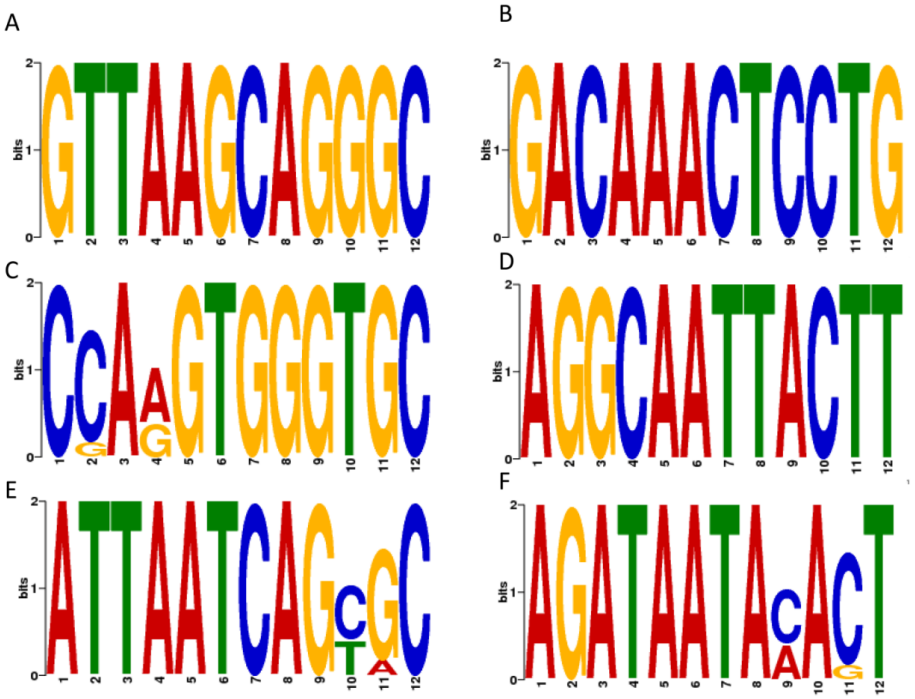

Figure 3. 22 Graphical representation of transcription factors binding motifs identified by MEME

Computationally predicted motifs in CNE1. Position frequency logos generated from human-teleost alignment. The bit score shows conservation in different lineages. Each of the motifs was further screened against the TRANSFAC library.

It is interesting to note that CNE1 is significantly expressed in blood precursor cells and circulatory red blood cells, it contains a tetrapod/teleost conserved putative binding site an important transcription factor GATA (Figure 3.23) known to be involved in the erythroid maturation during hematopoiesis (Galloway *et al.*, 2008). Similarly, binding site for Sox10 (GACAAACTCCTC) is present within CNE1 (Figure 3.23), and is involved in the vertebrate CNS development (Cheng *et al.*, 2000). Similarly, Table 3.2 shows putative transcription factors for other CNEs.

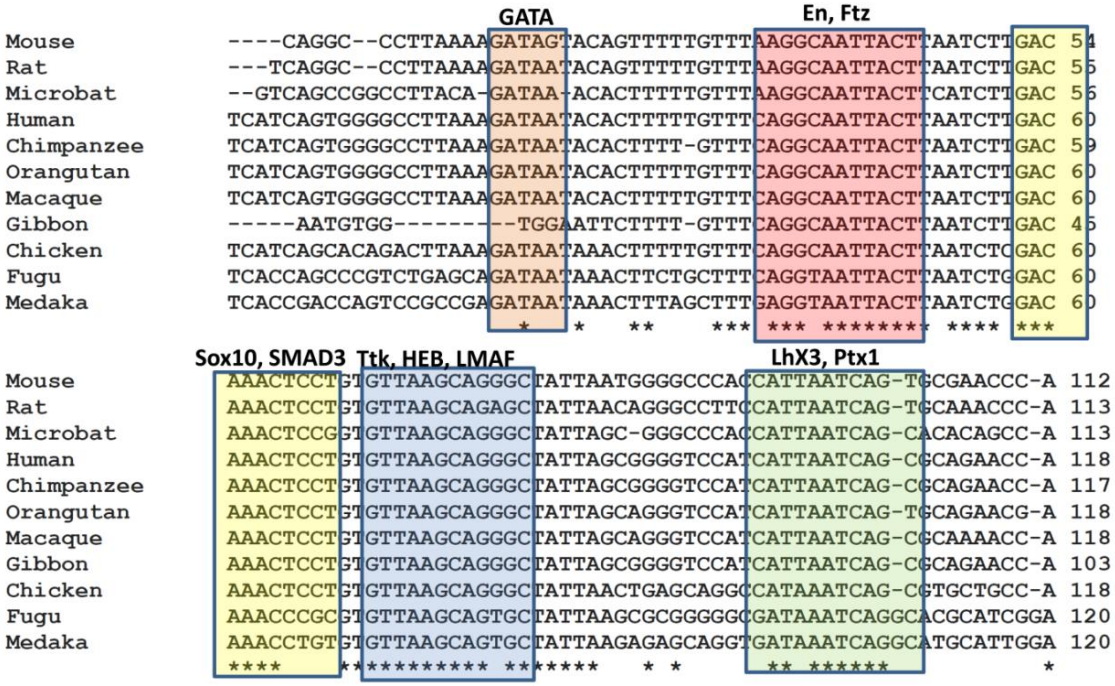

Figure 3. 23 Conserved transcription factor binding sites in CNE1.

Multiple alignment of human-teleost sequences with CLUSTALW. Spaces in between the alignment indicate deletion events, while a star symbol underneath represents a nucleotide position conserved in all lineages. Human/Fugu conserved putative binding sites for transcription factors are highlighted, and transcription regulators for conserved regions indicated above.

Table 3. 2 Computationally predicted transcription factors binding sites and their associated transcription factors

Element	Motif identified by MEME	Putative Transcription factors
CNE1	GTTAAGCAGGGC	Ttk, HEB/frem1
	GACAAACTCCTG	SMAD3, SOX10
	CCAAGTGGGTGC	Hmx3, Nkx2-5, Ralpha
	AGGCAATTACTT	En, Ftz, NF-kappaB
	ATTAATCAGCGC	Lhx3, OTX/Ptx1, NF-E2
	AGATAATACACT	GATA-3, GATA-2, Evi-1
CNE2	GATGAATGTGGT	CRP/Crip2, AML1/Runx1
	ATCCATCAAGCC	Pbx-1
	TGTCTAATTACA	HOXA4, CHX10/hox10, En
	TGCACAGCAAAT	TCF11/NFE2L1, OCT_4
	CCAGCACCAAAT	NRSF/REST, E2F/E2F1
	GTGCGGTGTCAT	TCF11/Nfe2l1, MAF, SREBP-1
CNE3	TATCAAGGGAAG	EBF, RORalpha1, TFII-I/Gtf2i
	AACAYATTTCCC	STAT1, HNF3-B/Foxa2,
	GGCGCATTAATA	OTX/Ptx1, POU3F2
	GGGRATAACAGA	Apex1, GATA-1/4, NF kappaB
	CTATTTAGTNAT	MEF-2, AGL15
CNE4	ATTTTCATGAAA	STAT5B, POU5F1, Oct_4
	CATGATGCATGG	Brn-2/Pou3f2, Pax-5
	TGATGACAAATT	CEBP, Cdc5, GATA-1
	TRAAATTGATGA	DMRT2, Pbx, En, OTX
	TTATTGTCAKGC	SOX17, Alx-4, Oct_1
CNE5	GCTAATAGAACC	AIRE, En
	ATGGATGGGCCC	RAP1, Pbx, NF-kappaB
	GCTAATGCTGTG	TCF11, Nrf-2
	TAACAATCTGGA	HOXA7, SOX17, ROX1
	GTTATTTACACA	FOXO3, FoxC1

DISCUSSION

Sequence comparisons of distantly related vertebrate species have revealed many genomic intervals that have remained conserved during vertebrate evolution (Abbasi *et al.*, 2007). Some of these sequences correspond to coding genes and non-coding RNAs, however two-thirds of them are unlikely to produce a functional transcript (Schweitzer *et al.*, 2000). These sequences fall in the new category of elements, which are termed as conserved non-coding elements (CNEs) (Woolfe *et al.*, 2005). Some of these elements are experimentally characterized to harbour transcription factor binding sites, and are implicated in the control of gene expression of various key developmental genes (Pennacchio *et al.*, 2006). Therefore, comparative genomics based strategies have emerged as a reliable methodology to predict genomic intervals harboring transcriptional regulatory elements even in the absence of knowledge about the specific characteristics of individual *cis*-regulatory elements (Nusslein-Volhard & Wieschaus, 1980).

The publication of comprehensive human genome is an enormous achievement but elucidating the entire grammar and encryption of functional elements encoded in the human and other closely related vertebrate genomes remains to be elucidated. The availability of different genome sequences in public databases and computational tools allows us to annotate such intervals which are highly conserved and present outside the coding regions. Comparative analysis of DNA sequences from distantly related species at varying evolutionary distances is a reliable approach for decoding the syntax of coding

and functional noncoding sequences, as well as sequences that are unique for a given lineage.

4.1 Regulation of Hh mediator is crucial for vertebrate embryogenesis

During vertebrate development, Hh proteins and their signaling pathways are fundamental for precise patterning and growth of several organs, including floor plate, neural tube, limb, and somites (Nieuwenhuis & Hui, 2005, Jiang & Hui, 2008, Yue *et al.*, 2009). Studies have shown that Hh proteins perform these functions by interacting with several downstream intracellular mediators (Hui & Angers, 2011, Jiang & Hui, 2008, Yue *et al.*, 2009). The transcriptional response to Shh is mediated by GLI1, GLI2 and GLI3. Gli^A function is reinforced by GLI1 which acts as transcriptional activator and a direct target of Hh signaling. The primary GLI^A activity is contributed by Gli2, whereas Gli^R function is largely contributed by Gli3 (Stecca & Altaba, 2010). The abnormality in the Hh pathway and its associated downstream mediators (GLI family) can influence the developmental events which ultimately lead to several human congenital malformations, including Gorlin syndrome, Greig cephalo polysyndactyly syndrome, and cancer including basal cell carcinoma and medulloblastoma (Nieuwenhuis & Hui, 2005). Many comprehensive studies have been done in revealing the role of GLI family in development and disease (Table 4.1). Despite the considerable progress in the study of Shh-Gli signaling pathway, the mechanism and molecular underpinnings that regulate the downstream effectors of this key signaling pathway still remains to be elucidated.

Table 4. 1 Summary of studies identifying GLI morphopathies

GLI1	***GLI2***	***GLI3***
Basal cell carcinoma Glioblastoma Osteosarcoma B-cell Lymphoma	Holoprosencephaly Defective anterior pituitary formation Branchial arch anomalies Pre/Post axial Polydactyly Brain and spinal cord defects (Mice) Foregut malformation Skin tumors Breast cancer Prostate cancer	GCPS PHS Preaxial polydactyly type IV (PPD-IV) and postaxial polydactyly type A (PAPA) Gli3-/- severe polydactyly and un-patterned digit Foregut malformation

4.2 Evolutionary sequence comparison reveals candidate *GLI2* enhancers

Studies have shown that systematically exploited non-coding conserved genomic intervals have functional potential and are clustered around developmental genes, which often act as tissue-specific enhancers (Woolfe & Elgar, 2008, Woolfe *et al.*, 2007). These genomic regions may be present within an intron, or upstream or downstream, and even in the intron of a neighboring gene. They can be identified by comparing the genomic sequences of distantly related species using several alignment tools (Nobrega *et al.*, 2003). Depending on the alignment tools, different levels of stringency have been applied for the identification of a manageable number of CNEs (putative enhancers), without a biologically based rationale. An obvious solution to highlight a functional

module in the genome is to compare the more distantly related species like human-fish, with suitable identification criteria.

Previously, our group has identified, using comparative analyses of human and fugu *GLI3* locus, 11 *GLI3*-associated CNEs distributed throughout the introns of *GLI3* gene, and reported them as tissue-specific enhancers (Abbasi *et al.*, 2007). The regulatory potential of a subset of *GLI3*-associated CNEs was in agreement with the reported Gli3/gli3 endogenous expression in vertebrates (Abbasi *et al.*, 2013, Abbasi *et al.*, 2007, Abbasi *et al.*, 2010). Similarly, several studies indicate that *GLI2* has essential functions controlling multiple patterning steps in different tissues/organs, and therefore a tight temporal and spatial control of gene expression is indispensable (Motoyama *et al.*, 2003, Motoyama *et al.*, 1998, Tojo *et al.*, 2003, McDermott *et al.*, 2005, Qi *et al.*, 2003). However, *cis*-regulatory underpinnings of the human GLI2 gene remains unknown. The identification of *cis*-acting regulatory elements interacting with the *GLI2* promoter could facilitate the detection of factors controlling the tissue-specific availability of GLI2 in *trans* in hedgehog target cells. In turn, identification of transcription factors for spatial and temporal control of GLI2 expression would greatly enhance our understanding of the regulatory network that coordinates the multitude of patterning events associated with the hedgehog-signaling pathway.

In the present study, *in silico* tools were employed to target the evolutionary conserved *cis*-regulatory elements of the human GLI2 gene encompassing 257kb length and having 14 exons. By employing multi-species sequence alignment using MLAGAN, an ancient

(tetrapod-teleost conserved) non-coding architecture was identified exclusively within the introns of *GLI2*. MLAGAN is a highly sophisticated tool to highlight the conserved genomic sequences in multiple species and it is based on the principle that orthologous regions have been identified between two or more species, and that there are no genomic rearrangements present within these regions. These human-fish ancient marks are surrounded by larger sequences preserved in evolutionarily more modern species such as mouse and chicken. The stringent criteria used to select these five CNEs were >50bp tracks with more than 50% sequence similarity between tetrapod and teleost. Comparative analysis of 100kb upstream or downstream region of the GLI2 gene did not detect any significant sequence similarity from human to fish. Moreover, none of the selected conserved non-coding elements overlap with exons or non-protein coding RNAs. The selected subset of intra-*GLI2* CNEs were BLAST against human genome using UCSC and Ensembl genome browsers, to verify whether these genomic intervals are unique to *GLI2* gene, or have a significant sequence similarity with the *GLI2* paralogs (*GLI1* and *GLI3*). This analysis revealed that all the five CNEs have significant hits against human *GLI2* gene only, and no evidence of overlap was found nearby or within the *GLI1* and *GLI3* genes. Similar to the previously identified *GLI3*-associated enhancers (Abbasi *et al.*, 2007), *GLI2* CNEs are distributed across almost the entire gene interval, remaining unchanged over the evolutionary distance of 450 Myr. CNE1 and CNE5 are present in intron 1, whereas highly conserved CNE2 is present within intron 2. CNE3 is located in intron 4, whereas CNE4 is present within intron 7 of *GLI2* gene. Together these studies

on *GLI* paralogs, suggest that these mediators of Hh pathway harbor conserved *cis*-regulatory signatures within their introns.

4.3 Progressive expansion of novel regulatory components around an ancient enhancer element

There are several studies which show that flanking intervals of the human-fish conserved sequence elements contribute to the activity of those elements and suggest that after the divergence of tetrapod-teleost lineages (450 Myr), there is a progressive gain-of-novel functions concerted around an ancient enhancer element (Abbasi *et al.*, 2007, Abbasi *et al.*, 2010, Poulin *et al.*, 2005). As *cis*-regulatory elements are extended in tetrapods, the selected subset of *GLI2* enhancers were defined such that they have the fish-specific core conserved track (necessary for the development of the basic structures) and flanking tetrapod-specific regions (required for the development of lineage-specific structures) (Abbasi *et al.*, 2010, Abbasi *et al.*, 2007). The strategy, used to define the boundaries of *GLI2* CNEs contain the functionally critical regulatory module (human-fish) as well as the flanking less conserved tetrapod-specific region. For example, within CNE1 (993bp), there is a highly conserved sequence track of ~129bp, almost 85% identical in human/rodents and with ~76% sequence identity in human/fugu. Table 3.1 summarizes the length of the selected subset of intra-*GLI2* CNEs, and their core conserved region between human to teleost.

4.4 *GLI2*-associated CNEs show tissue-specific regulatory activity *in vivo*

In order to address the *in vivo* role of *GLI2*-associated CNEs, I used transient reporter gene expression in zebrafish embryos with two different approaches: firstly, exploiting a co-injection strategy using a minimal β-globin promoter, and secondly, through direct cloning into a *Tol2* vector with a c-*fos* promoter (Abbasi *et al.*, 2007, Pauls *et al.*, 2012, Woolfe *et al.*, 2005). These approaches exploit the transparency and rapid development of zebrafish embryos, and have shown their potential for functionally testing enhancer elements among conserved non-coding regions (Fisher *et al.*, 2006, Woolfe *et al.*, 2005). GFP expression is observed in consistent pattern in co-injection assay, however, it is a laborious technique to inject and screen hundreds of embryos to generate a comprehensive view of the reporter expression pattern, and there is high mosaicism (Woolfe *et al.*, 2005). Moreover, failure of detecting reporter expression with co-injection assay for some *GLI2* intronic CNEs in fin illuminates the fact that this strategy is not suitable to detect the activity of limb-specific regulators. In comparison, *Tol2* transposon constructs injected with transposase mRNA integrates in the genome of somatic cells with high efficiency and tissue-specific GFP expression is observed easily due to non-mosaic reporter pattern and is a more suitable strategy for smaller domains like fin (Fisher *et al.*, 2006, Booker *et al.*, 2013). *Tol2* expression vector is derived from *Tol2* transposon, which were identified in the genome of *Oryzias latipes* (Medaka fish) (Koga *et al.*, 1996). Reporter gene expression depends on the activation of a *c-fos* minimal promoter, which can only induce reporter gene expression by inserting regulatory elements with positive activity. Several studies have already been reported to

support that this transient expression assay is useful for rapidly evaluating the non-coding genomic intervals (Pauls *et al.*, 2012, Fisher *et al.*, 2006).

Our results with two different approaches indicate that multiple evolutionarily conserved *GLI2*-associated human *cis*-regulators control highly coordinated GFP reporter gene expression in transient transgenic zebrafish, mimicking a subset of the known repertoire of endogenous *gli2* expression (Thisse & Thisse, 2005), and are apparently independent of the basal promoter used (Table 4.2 and 4.3). This functional study reveals the high propensity of conserved human non-coding sequences to behave as tissue-specific transcriptional enhancers *in vivo*, and support ancient human–fish conservation as highly effective filters to identify such functional elements. However, keeping the limitations of these approaches in mind about mosaicism (co-injection) and ectopic expression (*Tol2*), cautious analysis was done. Owing to the mosaic pattern of GFP reporters the expression pattern is not as evident as *in vivo* gli2 expression pattern. Moreover, reporter gene expression is sometimes detected in tissues where gli2 is not reported. Such ectopic expression may be caused by integration-position effects or high levels of presence of reporter constructs in several cells.

Reporter gene expression is observed in various domains coinciding with known sites of GLI2 activity (Table 4.3). For example, CNE1 drives GFP expression predominantly in various subdivisions of the CNS, forebrain, midbrain, hindbrain, and spinal cord (Table 4.2). The reporter gene activity was also frequent in the muscle fibers, ventral caudal region, skin and blood cells, consistent with the reported timing of zebrafish gli2

expression in these tissues (Thisse & Thisse, 2005) (Table 4.3). CNE2 activity was most frequent in notochord, brain, spinal cord, muscle fibers, blood cells and fin (Table 4.2). CNE3 activity was more specific to hindbrain, spinal cord, muscle cells, cardiac chamber, branchial arch and otic vesicles. CNE4 induced reporter gene expression in muscle cells. CNE5 activity was most frequent in the CNS and fin. Functional overlapping can be seen in a subset of enhancers with respect to the site of expression which is evident for all regulatory elements, a notion concordant with findings in other genes like GLI3 (Abbasi *et al.*, 2007, Abbasi *et al.*, 2013).

Table 4. 2 Comparison of reporter gene expression induced by intra-*GLI2* CNEs

Element	Forebrain	Midbrain	Hindbrain	Spinal cord	Notochord	Muscle cells	Fin	Otic vesicle	Cardiac cells	Epidermis
CNE1	+	+	+	+	–	+	–	–	+	+
CNE2	–	–	+	+	+	+	+	–	–	+
CNE3	–	–	+	+	–	+	–	+	+	–
CNE4	–	–	–	–	–	+	–	–		+
CNE5	–	–	+	+	–	+	+	–	+	–
Dr-*gli2a*_CNE2a	–	–	+	+	–	+	+	–	–	–
Dr-*gli2b*_CNE2b	–	–	+	–	–	+	+	–	–	–

Comparison of CNEs-mediated reporter gene expression in zebrafish embryonic domains. The "+" sign indicates the domain where GFP signal is detected, while "-" sign indicates those domains where GFP signal is not observed. CNE; Conserved non-coding element, Dr; Danio rerio, GFP; Green fluorescent protein.

Table 4. 3 Reported endogenous expression pattern of Gli2 and gli2a/gli2b in vertebrates

Gene	Endogenous expression	Source
GLI2/gli2a	Forebrain	(Hui et al., 1994)
	Midbrain	(Magdaleno et al., 2006), (Ke et al., 2005)
	Hindbrain	(Ke et al., 2008), (Thisse and Thisse, 2005), (Hui et al., 1994)
	Spinal cord	(Sasaki H, 1999), (Lee et al., 1997)
	Notochord	(Ding et al., 1998)
	Limb /Pectoral fin	(Mo et al., 1997), (Karlstrom et al., 2003)
	Muscle cells	(Du and Dienhart, 2001)
	Otic vesicles	(Hui et al., 1994)
gli2b	Forebrain	(Thisse and Thisse, 2005)
	Midbrain	(Ke et al., 2005)
	Hindbrain	(Ke et al., 2008)
	Spinal cord	(Thisse and Thisse, 2005)
	Notochord	(Thisse and Thisse, 2005)
	Pectoral fin	Need to be investigated
	Muscle cells	(Wang et al. 2013)
	Otic vesicles	(Ke et al., 2005)

This table provides information of already reported endogenous expression of Gli2 in mice and its co-orthologs (gli2a/gli2b) in zebrafish. The last column provides the source of literature that addresses the function and endogenous expression pattern of Gli2/gli2.

4.5 CNE2, CNE3, and CNE5 induce reporter expression that coincides with known sites of GLI2 activity in CNS

Gli2 plays an activator role in the hindbrain and spinal cord patterning; studies in $Gli2^{-/-}$ mice have shown diverse ventral patterning defects in hindbrain and spinal cord with a severely affected floor plate (FP) and interneurons (Ding *et al.*, 1998, Lebel *et al.*, 2007). Ke et al. have shown that during brain development in zebrafish *gli2b* is expressed in the rhombomere, whereas *gli2a* transcripts are expressed in the midbrain-hindbrain boundary (Ke *et al.*, 2008). Consistent with the role of GLi2/gli2s our data shows that reporter expression driven by a subset of identified *GLI2* intronic enhancers (CNE2, CNE3, and CNE5) is largely confined to the hindbrain and dorsal

spinal cord neurons, reflecting the complex role of *gli2* in neural tube development. These *in vivo* data revealed the overlapping contribution of CNE2, CNE3, and CNE5 in the precise patterning of the neural tube. Furthermore, *in silico* analysis of CNE2, CNE3, and CNE5 predicts human-fugu conserved binding sites for a number of developmentally important transcription factors like AML1, Pbx, HoxA, Foxa2, Ptx1, GATA1, Sox17 and Tcf11, that are known to be co-expressed with Gli2 during neural tube formation. Several studies show that Shh signalling is a prerequisite to control the patterning of the ventral neural tube at all rostro-caudal levels, from forebrain to spinal cord, through the activator function of Gli2 while Gli3 acts as a repressor in neural tube patterning (Persson *et al.*, 2002, Lebel *et al.*, 2007, Bai *et al.*, 2002). RNA *in situ* studies in zebrafish detected *gli2* in the anterior neural plate and as the development proceeds *gli2* can be detected uniformly throughout the dorsal forebrain, midbrain, hindbrain and is also expressed adjacent and dorsal to the cells expressing Shh (Karlstrom *et al.*, 1999, Ding *et al.*, 1998, Ruiz i Altaba, 1998). The hindbrain-specific expression territories of CNE2, CNE3, and CNE5 were further confirmed by testing these human genomic intervals in a transgenic zebrafish line that expresses RFP reporter gene in r3 (rhombomere3) and r5 (rhombomere5). These expression patterns are in agreement with endogenous *Gli2/gli2* expression in the CNS particularly in the hindbrain region, which is the most ancient part of the vertebrate brain (Table 4.3) (Jackman *et al.*, 2000, Ke *et al.*, 2008, Lebel *et al.*, 2007).

Intriguingly, it was observed that CNE2 only regulated reporter gene expression in the notochord at day-2 as well as at day-3 with the co-injection assay. However, using *Tol2* assay a very prominent expression was recapitulated in the notochord

and some other domains. Notochord has well-defined roles in patterning the surrounding tissues like neural tube and sends signals to develop the floor plate (Matise *et al.*, 1998). Among the signals secreted from notochord are the hedgehog proteins out of which the Shh ultimately defines the fate of ventral spinal cord (Yamada *et al.*, 1993, Holland *et al.*, 2004). In addition, mutant studies in mice show that GLI2 functions downstream of Shh and its activator function is required for the induction of the floor plate (Park *et al.*, 2000, Jacob & Briscoe, 2003). Shh establishes a gradient of Gli activity in the neural tube by inhibiting Gli repressor activity and potentiating its activator function. In *GLI2* mutant mice the notochord does not move away from the ventral spinal cord, as it does in wild type embryos (Ding *et al.*, 1998). RNA *in situ* studies in zebrafish have shown that gli2a and gli2b are widely expressed in the neural tube (Table 4.3). Reflecting the complex role of gli2 in neural tube, our results indicate that *GLI2*-associated CNE2 controls highly coordinated reporter gene expression in transgenic zebrafish, mimicking almost the entire known repertoire of endogenous Gli2 expression (Thisse & Thisse, 2005). Similarly, the core-conserved region between human and teleost of CNE2 was also investigated to study the role of flanking region around this highly conserved interval. Our findings show that the expression by truncated human CNE2 is almost similar to full length CNE2, except the pectoral fin where GFP expression is detected only with the full-length element. However, the reporter gene expression within the hindbrain and notochord was similar to full-length element. Recently several studies have shown that a limited number of sequence changes between conserved enhancers exhibiting more functional differences than similarities (Goode *et al.*, 2011).

4.6 *Cis*-regulatory control of GLI2 expression in developing pectoral fin

Gli2 expression patterns within the nascent limb bud are highly dynamic and context dependent. Establishment of anterior-posterior positional identities in the limb requires integration of the spatial distribution, timing, and dosage of GLI2 expression (Bowers *et al.*, 2012). RNA *in situ* studies in mice embryos have shown that at E10.5, Gli2 and Gli3 are broadly expressed in undifferentiated mesenchyme of the emerging limb bud, except posterior mesenchyme, where the genetic antagonism between Gli2 and Shh leads to reduced expression of Gli2 (Bai & Joyner, 2001, Mo *et al.*, 1997). $Gli2^{-/-}$ mice show reduced ossification in various bones in limbs including; stylopod (humerus, femur), zeugopod (radius, ulna, tibia and fibula) as well as shortening of the autopod (Mo *et al.*, 1997). In Gli3 mutant mice, Gli2 is vital for the patterning of anterior and posterior regions of the autopod (Bowers *et al.*, 2012). Moreover, studies in zebrafish showed that gli2a expression in the fin bud appears at 30 hpf (Thisse & Thisse, 2005), and by 36 hpf it is expressed uniformly throughout the mesenchyme of the developing pectoral fin (Karlstrom *et al.*, 2003). In addition, limb deformities like preaxial and postaxial polydactyly are well documented with *GLI2* mutations (Bertolacini *et al.*, 2012, Roessler *et al.*, 2005, Roessler *et al.*, 2003).

Consistent with a role for *gli2a* in fin, it was found that CNE2 and CNE5 drive reproducible reporter expression in the developing pectoral fin. The primary expansion of teleost-paired fins is parallel to that of tetrapod limb buds and is controlled by a similar mechanism (Zhang *et al.*, 2010). Evolution of *cis*-acting elements is proposed to be the key regulator in the development and structural

framework of the vertebrate fin/limb skeleton. Shh, Hox, hand2 and Gli family are the key signaling transduction pathways implicated in fin and limb morphologies (Abbasi, 2011, Mo *et al.*, 1997, Galli *et al.*, 2010, Tarchini & Duboule, 2006).

Overlapping activity of two independent enhancers in fin reflects the fact that *GLI2* harbors multiple *cis*-regulatory modules required for the normal development of limb. Furthermore, the roles of CNE2 and CNE5 in the pectoral fin are reflected by the presence of binding sites for numerous established transcription factors like HoxA, E2F, PTX1, and Nrf2, which are known to be co-expressed with Gli2 in the developing limb (http://www.informatics.jax.org/). Spatial and temporal activity of these enhancers needs to be investigated further in tetrapod model animals like mice, to completely decipher the mechanism of *Gli2* in anterior-posterior polarity of the limb and to define the expression pattern boundaries driven by *GLI2*-associated CNEs.

4.7 *Cis*-regulatory control of *GLI2* expression in skin cells

It is important to note that in addition to its critical role in neural tube, limb, and lung development, GLI2 plays an important role in skin cells, and its aberrant expression may lead to various carcinomas including BCC, an uncontrolled growth that arises in the skin's basal cells (Tojo *et al.*, 2003, Ding *et al.*, 1998, Mo *et al.*, 1997, Park *et al.*, 2000). Molecular and genetic studies have shown that Shh and Gli2 are required for BCC maintenance in a conditional expression mode (Hutchin *et al.*, 2005). Consistent with the GLI2 role and its endogenous expression in skin cells, three conserved intronic intervals, i.e. CNE1, CNE2, and CNE4, are able to mediate

the transcription of transgene in epidermis (Table 4.2). These findings reflect that *GLI2* intronic intervals are strong candidates for mutation screening of BCC patients which cannot be attributed to a mutation in the protein coding sequences of GLI2 gene.

4.8 CNE1, CNE3 and CNE5 activate reporter expression within cardiac chamber and circulatory blood cells

The Shh pathway participates in the establishment of cardiac progenitor cells during early heart development in zebrafish. Inhibition of the Shh signaling resulted in defect in myocardial progenitor specification leading to reduction of both ventricular and atrial cardiomyocytes. Moreover, activation of Shh signaling resulted in an increase of cardiomyocytes number (Voronova *et al.*, 2012, Thomas *et al.*, 2008). In addition, $Gli2^{-/-}Gli3^{+/-}$ mice indicated cardiac outflow tract anomalies (Kim *et al.*, 2001b). The importance of the Shh signaling pathway in mammalian heart development was demonstrated by total and tissue-specific knockout studies (Kim *et al.*, 2001a, Washington Smoak *et al.*, 2005). Thus, Shh signaling via *Gli2* is important for embryonic heart development. Consistent with role of GLI2 in cardiac chamber, three of the intronic enhancers (CNE1, CNE3, and CNE5) of GLI2 were observed to induce reporter gene expression in the cardiac cells. These anciently conserved genomic intervals might act together in a combinatorial fashion for development of cardiac cells.

4.9 CNE3 activity in branchial arch and otic vesicle

CNE3 drives GFP expression in branchial arch and otic vesicles after 24 and 48 hpf. The neural crest gives rise to the branchial arch derivatives of the craniofacial

skeleton. All three Gli genes are expressed in the neural crest derivatives of the craniofacial skeleton (Hui et al., 1994). Craniofacial anomalies involving the first branchial arch and the orbits have also been reported in patients with GLI2 mutations. Ke and colleagues have shown that cross-section of a 48 hpf embryo demonstrated that gli2b expression is present in two clusters in the medial part of the neural tube and bilaterally immediately above the otic vesicle, indicating a possible relationship between these lateral clusters and ear function (Ke *et al.*, 2005). Although there is no direct evidence of gli2 role in the development and its expression is not so strong in zebrafish ear, but Hh signaling is necessary for accurate AP patterning of the zebrafish otic vesicle (Hammond *et al.*, 2003). Interestingly, CNE3 also induced GFP expression with low frequency within cranial ganglia in zebrafish embryos at day-3. Gli2 expression in mouse was not detected in cranial ganglion and is not so far described in zebrafish. Thus, the expression of gli2 in zebrafish might be more widespread than reported so far.

4.10 *GLI2*-associated CNEs appear to drive overlapping GFP expression in zebrafish muscle

Interestingly, almost all the human CNEs and zebrafish CNEs (CNE2a and CNE2b) drive reporter gene expression in muscle either with co-injection or *Tol2*, or with both the assays. Zebrafish gli2 has been shown to express in fast and slow muscle precursor cells in the segmental plate and developing somites (Du & Dienhart, 2001, Wang *et al.*, 2013). Together these data, suggest key roles of Gli2 in the normal induction of muscles in zebrafish.

4.11 Duplicated CNEs suggest overlapping expression pattern in hindbrain and pectoral fin

In teleost, *Gli2* underwent duplication producing two gene copies, namely *gli2a* and *gli2b* (Ke *et al.*, 2005). It has already been reported that combined activity of *gli2a* and *gli2b* of zebrafish play a crucial role in Hh signaling during embryogenesis (Ke *et al.*, 2008). Interestingly, comparative sequence analysis revealed two copies (co-orthologs) of human CNE2 in zebrafish. One of the copies (CNE2a) was positioned within intron 2 of the *gli2a* gene whilst the other copy (CNE2b) resides downstream of the *gli2b* gene. CNE2a (119bp) and CNE2b (117bp) share significant sequence similarity with each other (75%), and also with human *GLI2*-CNE2 (117bp), i.e. 68% and 65% respectively.

I analyzed each of these elements independently in transgenic zebrafish and compared reporter gene expression induced by these elements with each other. Careful examination of resultant data shows apparently overlapping reporter expression of CNE2a and CNE2b in zebrafish hindbrain (Fig. 7A-7D) and developing fin (Fig. 7E and 7F), wherein endogenous expression of gli2a/gli2b genes in a complimentary manner has already been reported (Table 3) (Thisse and Thisse, 2005; Ke et al., 2008). To decipher the reason for this overlapping expression of the duplicated CNEs, and any indication of subfunctionalization involved in the process, more precise assays need to be performed. These findings reflect that the combined activity of gli2a/gli2b *cis*-regions is essential for the normal development of zebrafish hindbrain and fin.

In addition to hindbrain and fin, CNE2a induced GFP expression in the spinal cord, circulating blood cells and muscle cells. As compared to CNE2a, CNE2b does not activate GFP signal enough in the spinal cord, however the reporter gene expression in the circulating blood and muscle cells is similar by dCNEs. Thus, there is a partial overlap and divergence in the activities of these two enhancers during early embryogenesis.

4.12 Conclusions and future perspectives

In the past decade, identification of conserved non-coding elements became the bench mark for the search for functional modules across model organisms. Zebrafish has been shown as a good model for the characterization of components of Hh signaling including GLI family members which are key transcription factors in vertebrate development. The regulation of GLI family evolved and fine-tuned rapidly from their ancient homolog *Ci*, which functions primarily within a context of the transduction of Hh signaling. GLI1 and GLI2 mainly act as an activator and GLI3 performs repressor function. The presence of three paralogs in vertebrate genome with different functions within and between the species compel to study the difference of regulation between these closely related family members. Understanding the gene regulation of *GLI2*, an important GLI family member, is essential as it helps in deciphering the mechanism of how this key gene works. Elucidating the regulation of *GLI2* expression is as difficult as understanding the function of this transcription factor, since to date, no *cis*-acting control of human *GLI2* gene was known.

In this study, for the first time, *cis*-acting regulatory control associated with the human GLI2 gene was explored and studied in zebrafish using transient transgenic assays. This data revealed that *GLI2* gene harbors deeply conserved genomic intervals that control the expression of the GLI2 transcript during early embryogenesis. Their expression pattern suggests that *GLI2*-associated intronic enhancers play an overlapping role in several key developmental domains and consistent with the role of GLI2/gli2 in vertebrates.

Further studies can be done to test these *GLI2*-associated enhancers in modern tetrapod model animals like chicken, frog or mice, to validate the role of these elements during embryogenesis, using orthologous sequences. This strategy would help to redefine the mechanism of *GLI2* gene regulation in its genomic context. Studies involving deletion analysis of TFBSs within each enhancer would help in an in depth analysis of their critical regions.

Taken together, the data presented in this thesis, will not only help to better understand the genetic mechanism of Shh-GLI signaling but also proposes to look for the role of this catalogue in evolutionary events involved in the structural anatomy and functional diversification of vertebrates. In addition, these *cis*-regulatory modules can fill the gap for mutational studies of *GLI2*-associated human birth defects, which cannot be attributed to any exonic mutations.

REFERENCES

Abbasi, A. A. 2011. Evolution of vertebrate appendicular structures: Insight from genetic and palaeontological data. *Dev Dyn,* **240,** 1005-1016.

Abbasi, A. A., Goode, D. K., Amir, S. & Grzeschik, K. H. 2009. Evolution and functional diversification of the GLI family of transcription factors in vertebrates. *Evol Bioinform Online,* **5,** 5-13.

Abbasi, A. A., Minhas, R., Schmidt, A., Koch, S. & Grzeschik, K. H. 2013. Cis-regulatory underpinnings of human GLI3 expression in embryonic craniofacial structures and internal organs. *Dev Growth Differ,* **55,** 699-709.

Abbasi, A. A., Paparidis, Z., Malik, S., Bangs, F., Schmidt, A., Koch, S., Lopez-Rios, J. & Grzeschik, K. H. 2010. Human intronic enhancers control distinct sub-domains of Gli3 expression during mouse CNS and limb development. *BMC Dev Biol,* **10,** 44.

Abbasi, A. A., Paparidis, Z., Malik, S., Goode, D. K., Callaway, H., Elgar, G. & Grzeschik, K. H. 2007. Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers. *PLoS One,* **2,** e366.

Alcedo, J., Ayzenzon, M., Von Ohlen, T., Noll, M. & Hooper, J. E. 1996. The Drosophila smoothened gene encodes a seven-pass membrane protein, a putative receptor for the hedgehog signal. *Cell,* **86,** 221-232.

Altaba, R. I. 1999. Gli proteins and Hedgehog signaling: development and cancer. *Trends in Genetics,* **15,** 418425.

Alvarez-Medina, R., Cayuso, J., Okubo, T., Takada, S. & Marti, E. 2008. Wnt canonical pathway restricts graded Shh/Gli patterning activity through the regulation of Gli3 expression. *Development,* **135,** 237-247.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H. & Shiroishi, T. 2009. Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev Cell,* **16,** 47-57.

Anelli, V., Santoriello, C., Distel, M., Koster, R. W., Ciccarelli, F. D. & Mione, M. 2009. Global repression of cancer gene expression in a zebrafish model of melanoma is linked to epigenetic regulation. *Zebrafish,* **6,** 417-424.

Bai, C. B., Auerbach, W., Lee, J. S., Stephen, D. & Joyner, A. L. 2002. Gli2, but not Gli1, is required for initial Shh signaling and ectopic activation of the Shh pathway. *Development,* **129,** 4753-4761.

Bai, C. B. & Joyner, A. L. 2001. Gli1 can rescue the in vivo function of Gli2. *Development,* **128,** 5161-5172.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res,* **37,** W202-208.

Bertolacini, C. D., Ribeiro-Bicudo, L. A., Petrin, A., Richieri-Costa, A. & Murray, J. C. 2012. Clinical findings in patients with GLI2 mutations--phenotypic variability. *Clin Genet,* **81,** 70-75.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E. *et al.* 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* **447,** 799-816.

Bolker, J. A. 2000. Modularity in Development and Why It Matters to Evo-Devo. *Integrative and Comparative Biology,* **40**.

Booker, B. M., Murphy, K. K. & Ahituv, N. 2013. Functional analysis of limb enhancers in the developing fin. *Development genes and evolution,* **223,** 395-399.

Bowers, M., Eng, L., Lao, Z., Turnbull, R. K., Bao, X., Riedel, E., Mackem, S. & Joyner, A. L. 2012. Limb anterior-posterior polarity integrates activator and repressor functions of GLI2 as well as GLI3. *Dev Biol,* **370,** 110-124.

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A. & Batzoglou, S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res,* **13,** 721-731.

Butler, J. E. & Kadonaga, J. T. 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev,* **16,** 2583-2592.

Carroll , S. 2005. Evolution at Two Levels: On Genes and Form. *PLoS Biology,* **3**.

Carroll, S. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell,* **134,** 25-36.

Chakalova, L., Carter, D., Debrand, E., Goyenechea, B., Horton, A., Miles, J., Osborne, C. & Fraser, P. 2005. Developmental regulation of the beta-globin gene locus. *Prog Mol Subcell Biol,* **38,** 183-206.

Chen, L. & Widom, J. 2005. Mechanism of transcriptional silencing in yeast. *Cell,* **120,** 37-48.

Cheng, W. T., Xu, K., Tian, D. Y., Zhang, Z. G., Liu, L. J. & Chen, Y. 2009. Role of Hedgehog signaling pathway in proliferation and invasiveness of hepatocellular carcinoma cells. *Int J Oncol,* **34,** 829-836.

Cheng, Y., Cheung, M., Abu-Elmagd, M. M., Orme, A. & Scotting, P. J. 2000. Chick sox10, a transcription factor expressed in both early neural crest cells and central nervous system. *Brain Res Dev Brain Res,* **121,** 233-241.

Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L. & Myers, R. M. 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res,* **16,** 1-10.

Coy, S., Caamano, J. H., Carvajal, J., Cleary, M. L. & Borycki, A. G. 2011. A novel Gli3 enhancer controls the Gli3 spatiotemporal expression pattern through a TALE homeodomain protein binding site. *Mol Cell Biol,* **31,** 1432-1443.

Dahmane, N., Lee, J., Robins, P., Heller, P. & Ruiz I Altaba, A. 1997. Activation of the transcription factor Gli1 and the Sonic hedgehog signalling pathway in skin tumours. *Nature,* **389,** 876-881.

Dai P, A. H., Tanaka Y, Maekawa T, Nakafuku M, Ishii S. 1999. Sonic Hedgehog-induced activation of the Gli1 promoter is mediated by GLI3. *J Biol Chem.,* **274**.

Denef, N., Neubuser, D., Perez, L. & Cohen, S. M. 2000. Hedgehog induces opposite changes in turnover and subcellular localization of patched and smoothened. *Cell,* **102,** 521-531.

Ding, Q., Motoyama, J., Gasca, S., Mo, R., Sasaki, H., Rossant, J. & Hui, C. C. 1998. Diminished Sonic hedgehog signaling and lack of floor plate differentiation in Gli2 mutant mice. *Development,* **125,** 2533-2543.

Distel, M., Wullimann, M. F. & Koster, R. W. 2009. Optimized Gal4 genetics for permanent gene expression mapping in zebrafish. *Proc Natl Acad Sci U S A,* **106,** 13365-13370.

Du, S. J. & Dienhart, M. 2001. Gli2 mediation of hedgehog signals in slow muscle induction in zebrafish. *Differentiation,* **67,** 84-91.

Durick, K., Mendlein, J. & Xanthopoulos, K. G. 1999. Hunting with traps: genome-wide strategies for gene discovery and functional analysis. *Genome Res,* **9,** 1019-1025.

Elgar, G. & Vavouri, T. 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet,* **24,** 344-352.

Ellingsen, S., Laplante, M. A., Konig, M., Kikuta, H., Furmanek, T., Hoivik, E. A. & Becker, T. S. 2005. Large-scale enhancer detection in the zebrafish genome. *Development,* **132,** 3799-3811.

Elson, E., Perveen, R., Donnai, D., Wall, S. & Black, G. C. 2002. De novo GLI3 mutation in acrocallosal syndrome: broadening the phenotypic spectrum of GLI3 defects and overlap with murine models. *J Med Genet,* **39,** 804-806.

Epstein, D. J. 2009. Cis-regulatory mutations in human disease. *Brief Funct Genomic Proteomic,* **8,** 310-316.

Felsenfeld, G., Burgess-Beusse, B., Farrell, C., Gaszner, M., Ghirlando, R., Huang, S., Jin, C., Litt, M., Magdinier, F., Mutskov, V. *et al.* 2004. Chromatin boundaries and chromatin domains. *Cold Spring Harb Symp Quant Biol,* **69,** 245-250.

Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L., Urasaki, A., Kawakami, K. & Mccallion, A. S. 2006. Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nat Protoc,* **1,** 1297-1305.

Franca, M. M., Jorge, A. A., Carvalho, L. R., Costalonga, E. F., Otto, A. P., Correa, F. A., Mendonca, B. B. & Arnhold, I. J. 2013. Relatively high frequency of non-synonymous GLI2 variants in patients with congenital hypopituitarism without holoprosencephaly. *Clin Endocrinol (Oxf),* **78,** 551-557.

Fulda, S., Meyer, E. & Debatin, K. M. 2002. Inhibition of TRAIL-induced apoptosis by Bcl-2 overexpression. *Oncogene,* **21,** 2283-2294.

Galli, A., Robay, D., Osterwalder, M., Bao, X., Benazet, J. D., Tariq, M., Paro, R., Mackem, S. & Zeller, R. 2010. Distinct roles of Hand2 in initiating polarity and posterior Shh expression during the onset of mouse limb bud development. *PLoS Genet,* **6,** e1000901.

Galloway, J. L., Wingert, R. A., Thisse, C., Thisse, B. & Zon, L. I. 2008. Combinatorial regulation of novel erythroid gene expression in zebrafish. *Exp Hematol,* **36,** 424-432.

Gaszner, M. & Felsenfeld, G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet,* **7,** 703-713.

Ghali, L., Wong, S. T., Green, J., Tidman, N. & Quinn, A. G. 1999. Gli1 protein is expressed in basal cell carcinomas, outer root sheath keratinocytes and a subpopulation of mesenchymal cells in normal human skin. *J Invest Dermatol,* **113,** 595-599.

Goode, D. K., Callaway, H. A., Cerda, G. A., Lewis, K. E. & Elgar, G. 2011. Minor change, major difference: divergent functions of highly conserved cis-regulatory elements subsequent to whole genome duplication events. *Development (Cambridge, England),* **138,** 879-884.

Grachtchouk M, M. R., Yu S, Zhang X, Sasaki H, Hui Cc, Dlugosz Aa. 2000. Basal cell carcinomas in mice overexpressing Gli2 in skin. *Nat Genet.,* **24,** 216-217.

Hammond, K., Loynes, H., Folarin, A., Smith, J. & Whitfield, T. 2003. Hedgehog signalling is required for correct anteroposterior patterning of the zebrafish otic vesicle. *Development (Cambridge, England),* **130,** 1403-1417.

Hardcastle, Z., Mo, R., Hui, C. C. & Sharpe, P. T. 1998. The Shh signalling pathway in tooth development: defects in Gli2 and Gli3 mutants. *Development,* **125,** 2803-2811.

Hardison, R., Slightom, J. L., Gumucio, D. L., Goodman, M., Stojanovic, N. & Miller, W. 1997. Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. *Gene,* **205,** 73-94.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A. & Saint-André, V. 2013. Super-enhancers in the control of cell identity and disease. *Cell*.

Holland, L. Z., Laudet, V. & Schubert, M. 2004. The chordate amphioxus: an emerging model organism for developmental biology. *Cell Mol Life Sci,* **61,** 2290-2308.

Hui, C. C. & Angers, S. 2011. Gli proteins in development and disease. *Annu Rev Cell Dev Biol,* **27,** 513-537.

Hui, C. C. & Joyner, A. L. 1993. A mouse model of greig cephalopolysyndactyly syndrome: the extra-toesJ mutation contains an intragenic deletion of the Gli3 gene. *Nat Genet,* **3,** 241-246.

Hui, C. C., Slusarski, D., Platt, K. A., Holmgren, R. & Joyner, A. L. 1994. Expression of three mouse homologs of the Drosophila segment polarity gene cubitus interruptus, Gli, Gli-2, and Gli-3, in ectoderm- and mesoderm-derived tissues suggests multiple roles during postimplantation development. *Dev Biol,* **162,** 402-413.

Hutchin, M. E., Kariapper, M. S., Grachtchouk, M., Wang, A., Wei, L., Cummings, D., Liu, J., Michael, L. E., Glick, A. & Dlugosz, A. A. 2005. Sustained Hedgehog signaling is required for basal cell carcinoma proliferation and survival: conditional skin tumorigenesis recapitulates the hair growth cycle. *Genes Dev,* **19,** 214-223.

Ikram, M. S., Neill, G. W., Regl, G., Eichberger, T., Frischauf, A. M., Aberger, F., Quinn, A. & Philpott, M. 2004. GLI2 is expressed in normal human epidermis and BCC and induces GLI1 expression by binding to its promoter. *J Invest Dermatol,* **122,** 1503-1509.

Jackman, W. R., Langeland, J. A. & Kimmel, C. B. 2000. islet reveals segmentation in the Amphioxus hindbrain homolog. *Dev Biol,* **220,** 16-26.

Jacob, J. & Briscoe, J. 2003. Gli proteins and the control of spinal-cord patterning. *EMBO Rep,* **4,** 761-765.

Jean-Jack, M. J. R. 2010. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods in molecular biology (Clifton, N.J.),* **674,** 33-42.

Jiang, J. & Hui, C. C. 2008. Hedgehog Signaling in Development and Cancer. *Dev Cell.,* **15,** 801-812.

Johnston, J. J., Sapp, J. C., Turner, J. T., Amor, D., Aftimos, S., Aleck, K. A., Bocian, M., Bodurtha, J. N., Cox, G. F., Curry, C. J. *et al.* 2010. Molecular analysis expands the spectrum of phenotypes associated with GLI3 mutations. *Hum Mutat,* **31,** 1142-1154.

Juan, A. H. & Ruddle, F. H. 2003. Enhancer timing of Hox gene expression: deletion of the endogenous Hoxc8 early enhancer. *Development,* **130,** 4823-4834.

Kang, S., Graham, J. M., Jr., Olney, A. H. & Biesecker, L. G. 1997. GLI3 frameshift mutations cause autosomal dominant Pallister-Hall syndrome. *Nat Genet,* **15,** 266-268.

Karlstrom, R. O., Talbot, W. S. & Schier, A. F. 1999. Comparative synteny cloning of zebrafish you-too: mutations in the Hedgehog target gli2 affect ventral forebrain patterning. *Genes Dev,* **13,** 388-393.

Karlstrom, R. O., Tyurina, O. V., Kawakami, A., Nishioka, N., Talbot, W. S., Sasaki, H. & Schier, A. F. 2003. Genetic analysis of zebrafish gli1 and gli2 reveals divergent requirements for gli genes in vertebrate development. *Development,* **130,** 1549-1564.

Kawakami, K., Takeda, H., Kawakami, N., Kobayashi, M., Matsuda, N. & Mishina, M. 2004. A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev Cell,* **7,** 133-144.

Ke, Z., Emelyanov, A., Lim, S. E., Korzh, V. & Gong, Z. 2005. Expression of a novel zebrafish zinc finger gene, gli2b, is affected in Hedgehog and Notch signaling related mutants during embryonic development. *Dev Dyn,* **232,** 479-486.

Ke, Z., Kondrichin, I., Gong, Z. & Korzh, V. 2008. Combined activity of the two Gli2 genes of zebrafish play a major role in Hedgehog signaling during zebrafish neurodevelopment. *Mol Cell Neurosci,* **37,** 388-401.

Kim, J., Kim, P. & Hui, C. C. 2001a. The VACTERL association: lessons from the Sonic hedgehog pathway. *Clin Genet,* **59,** 306-315.

Kim, P. C., Mo, R. & Hui Cc, C. 2001b. Murine models of VACTERL syndrome: Role of sonic hedgehog signaling pathway. *J Pediatr Surg,* **36,** 381-384.

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. 1995. Stages of embryonic development of the zebrafish. *Dev Dyn,* **203,** 253-310.

Kinzler, K. W., Ruppert, J. M., Bigner, S. H. & Vogelstein, B. 1988. The GLI gene is a member of the Kruppel family of zinc finger proteins. *Nature,* **332,** 371-374.

Kleinjan, D. J. & Coutinho, P. 2009. Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. *Brief Funct Genomic Proteomic,* **8,** 317-332.

Koga, A., Suzuki, M., Inagaki, H., Bessho, Y. & Hori, H. 1996. Transposable element in fish. *Nature,* **383,** 30.

Kurokawa, D., Kiyonari, H., Nakayama, R., Kimura-Yoshida, C., Matsuo, I. & Aizawa, S. 2004. Regulation of Otx2 expression and its functions in mouse forebrain and midbrain. *Development,* **131,** 3319-3331.

Lebel, M., Mo, R., Shimamura, K. & Hui, C. C. 2007. Gli2 and Gli3 play distinct roles in the dorsoventral patterning of the mouse hindbrain. *Dev Biol,* **302,** 345-355.

Lee, J., Platt, K. A., Censullo, P. & Ruiz I Altaba, A. 1997. Gli1 is a target of Sonic hedgehog that induces ventral neural tube development. *Development,* **124,** 2537-2552.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., De Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E. & De Graaff, E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet,* **12,** 1725-1735.

Levine, M. 2010. Transcriptional enhancers in animal development and evolution. *Current biology : CB,* **20,** 63.

Levine, M. & Tjian, R. 2003. Transcription regulation and animal diversity. *Nature,* **424,** 147-151.

Liu, C. Z., Yang, J. T., Yoon, J. W., Villavicencio, E., Pfendler, K., Walterhouse, D. & Iannaccone, P. 1998. Characterization of the promoter region and genomic organization of GLI, a member of the Sonic hedgehog-Patched signaling pathway. *Gene,* **209,** 1-11.

Liu, J., Prickett, T. D., Elliott, E., Meroni, G. & Brautigan, D. L. 2001. Phosphorylation and microtubule association of the Opitz syndrome protein mid-1 is regulated by protein phosphatase 2A via binding to the regulatory subunit alpha 4. *Proc Natl Acad Sci U S A,* **98,** 6650-6655.

Lomvardas, S., Barnea, G., Pisapia, D. J., Mendelsohn, M., Kirkland, J. & Axel, R. 2006. Interchromosomal interactions and olfactory receptor choice. *Cell,* **126,** 403-413.

Loots, G. G. & Ovcharenko, I. 2004. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res,* **32,** W217-221.

Loven, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I. & Young, R. A. 2013. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell,* **153,** 320-334.

Mager, U., Kolehmainen, M., De Mello, V. D., Schwab, U., Laaksonen, D. E., Rauramaa, R., Gylling, H., Atalay, M., Pulkkinen, L. & Uusitupa, M. 2008. Expression of ghrelin gene in peripheral blood mononuclear cells and plasma ghrelin concentrations in patients with metabolic syndrome. *Eur J Endocrinol,* **158,** 499-510.

Mahony, S. & Benos, P. V. 2007. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res,* **35,** W253-258.

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

Matise, M. P., Epstein, D. J., Park, H. L., Platt, K. A. & Joyner, A. L. 1998. Gli2 is required for induction of floor plate and adjacent cells, but not most ventral neurons in the mouse central nervous system. *Development,* **125,** 2759-2770.

Matys, V., Fricke, E., Geffers, 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.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S. & Dubchak, I. 2000. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics,* **16,** 1046-1047.

Mcdermott, A., Gustafsson, M., Elsam, T., Hui, C. C., Emerson, C. P., Jr. & Borycki, A. G. 2005. Gli2 and Gli3 have redundant and context-dependent function in skeletal muscle formation. *Development,* **132,** 345-357.

Mcewen, G. K., Woolfe, A., Goode, D., Vavouri, T., Callaway, H. & Elgar, G. 2006. Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. *Genome Res,* **16,** 451-465.

Methot, N. & Basler, K. 2001. An absolute requirement for Cubitus interruptus in Hedgehog signaling. *Development,* **128,** 733-742.

Mo, R., Freer, A. M., Zinyk, D. L., Crackower, M. A., Michaud, J., Heng, H. H., Chik, K. W., Shi, X. M., Tsui, L. C., Cheng, S. H. *et al.* 1997. Specific and redundant

functions of Gli2 and Gli3 zinc finger genes in skeletal patterning and development. *Development,* **124,** 113-123.

Motoyama, J., Liu, J., Mo, R., Ding, Q., Post, M. & Hui, C. C. 1998. Essential function of Gli2 and Gli3 in the formation of lung, trachea and oesophagus. *Nat Genet,* **20,** 54-57.

Motoyama J, L. J., Mo R, Ding Q, Post M, Hui Cc. 1998. Essential function of Gli2 and Gli3 in the formation of lung, trachea and oesophagus. *Nat Genet. ,* **20,** 54-57.

Motoyama, J., Milenkovic, L., Iwama, M., Shikata, Y., Scott, M. P. & Hui, C. C. 2003. Differential requirement for Gli2 and Gli3 in ventral neural cell fate specification. *Dev Biol,* **259,** 150-161.

Muller, F., Chang, B., Albert, S., Fischer, N., Tora, L. & Strahle, U. 1999. Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. *Development,* **126,** 2103-2116.

Nieuwenhuis, E. & Hui, C. C. 2005. Hedgehog signaling and congenital malformations. *Clinical genetics,* **67,** 193-208.

Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. 2003. Scanning human gene deserts for long-range enhancers. *Science,* **302,** 413.

Nusslein-Volhard, C. & Wieschaus, E. 1980. Mutations affecting segment number and polarity in Drosophila. *Nature,* **287,** 795-801.

Okano, H., Shiraki, K., Inoue, H., Kawakita, T., Yamanaka, T., Deguchi, M., Sugimoto, K., Sakai, T., Ohmori, S., Fujikawa, K. *et al.* 2003. Cellular FLICE/caspase-8-inhibitory protein as a principal regulator of cell death and survival in human hepatocellular carcinoma. *Lab Invest,* **83,** 1033-1043.

Osorio, A., Milne, R. L., Pita, G., Peterlongo, P., Heikkinen, T., Simard, J., Chenevix-Trench, G., Spurdle, A. B., Beesley, J., Chen, X. *et al.* 2009. Evaluation of a candidate breast cancer associated SNP in ERCC4 as a risk modifier in BRCA1 and BRCA2 mutation carriers. Results from the Consortium of Investigators of Modifiers of BRCA1/BRCA2 (CIMBA). *Br J Cancer,* **101,** 2048-2054.

Paparidis, Z., Abbasi, A. A., Malik, S., Goode, D. K., Callaway, H., Elgar, G., Degraaff, E., Lopez-Rios, J., Zeller, R. & Grzeschik, K. H. 2007. Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. *Dev Growth Differ,* **49,** 543-553.

Paquet, D., Bhat, R., Sydow, A., Mandelkow, E. M., Berg, S., Hellberg, S., Falting, J., Distel, M., Koster, R. W., Schmid, B. *et al.* 2009. A zebrafish model of tauopathy allows in vivo imaging of neuronal cell death and drug evaluation. *J Clin Invest,* **119,** 1382-1395.

Park, H. L., Bai, C., Platt, K. A., Matise, M. P., Beeghly, A., Hui, C. C., Nakashima, M. & Joyner, A. L. 2000. Mouse Gli1 mutants are viable but have defects in SHH signaling in combination with a Gli2 mutation. *Development,* **127,** 1593-1605.

Parker, H. J., Piccinelli, P., Sauka-Spengler, T., Bronner, M. & Elgar, G. 2011. Ancient Pbx-Hox signatures define hundreds of vertebrate developmental enhancers. *BMC Genomics,* **12,** 637.

Pauls, S., Smith, S. F. & Elgar, G. 2012. Lens development depends on a pair of highly conserved Sox21 regulatory elements. *Dev Biol,* **365,** 310-318.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D. *et al.* 2006. In vivo

enhancer analysis of human conserved non-coding sequences. *Nature,* **444,** 499-502.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. 2013. Enhancers: five essential questions. *Nat Rev Genet,* **14,** 288-295.

Pennacchio, L. A. & Rubin, E. M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet,* **2,** 100-109.

Persson, M., Stamataki, D., Te Welscher, P., Andersson, E., Bose, J., Ruther, U., Ericson, J. & Briscoe, J. 2002. Dorsal-ventral patterning of the spinal cord requires Gli3 transcriptional repressor activity. *Genes Dev,* **16,** 2865-2878.

Philipp, M. & Caron, M. G. 2009. Hedgehog signaling: is Smo a G protein-coupled receptor? *Current Biology,* **19**.

Pillai, S. & Chellappan, S. P. 2009. ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. *Methods Mol Biol,* **523,** 341-366.

Pohl, T. M., Mattei, M. G. & Ruther, U. 1990. Evidence for allelism of the recessive insertional mutation add and the dominant mouse mutation extra-toes (Xt). *Development,* **110,** 1153-1157.

Postlethwait, J. H., Woods, I. G., Ngo-Hazelett, P., Yan, Y. L., Kelly, P. D., Chu, F., Huang, H., Hill-Force, A. & Talbot, W. S. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Res,* **10,** 1890-1902.

Poulin, F., Nobrega, M. A., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E. M. & Pennacchio, L. A. 2005. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics,* **85,** 774-781.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E. M., Couronne, O. & Pennacchio, L. A. 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res,* **16,** 855-863.

Qi, Y., Tan, M., Hui, C. C. & Qiu, M. 2003. Gli2 is required for normal Shh signaling and oligodendrocyte development in the spinal cord. *Mol Cell Neurosci,* **23,** 440-450.

Radhakrishna, U., Bornholdt, D., Scott, H. S., Patel, U. C., Rossier, C., Engel, H., Bottani, A., Chandal, D., Blouin, J. L., Solanki, J. V. *et al.* 1999. The phenotypic spectrum of GLI3 morphopathies includes autosomal dominant preaxial polydactyly type-IV and postaxial polydactyly type-A/B; No phenotype prediction from the position of GLI3 mutations. *Am J Hum Genet,* **65,** 645-655.

Radhakrishna, U., Wild, A., Grzeschik, K. H. & Antonarakis, S. E. 1997. Mutation in GLI3 in postaxial polydactyly type A. *Nat Genet,* **17,** 269-271.

Rahimov, F., Ribeiro, L. A., De Miranda, E., Richieri-Costa, A. & Murray, J. C. 2006. GLI2 mutations in four Brazilian patients: how wide is the phenotypic spectrum? *Am J Med Genet A,* **140,** 2571-2576.

Roberts, W. M., Douglass, E. C., Peiper, S. C., Houghton, P. J. & Look, A. T. 1989. Amplification of the gli gene in childhood sarcomas. *Cancer Res,* **49,** 5407-5413.

Roessler, E., Du, Y. Z., Mullor, J. L., Casas, E., Allen, W. P., Gillessen-Kaesbach, G., Roeder, E. R., Ming, J. E., Ruiz I Altaba, A. & Muenke, M. 2003. Loss-of-function mutations in the human GLI2 gene are associated with pituitary

anomalies and holoprosencephaly-like features. *Proc Natl Acad Sci U S A,* **100,** 13424-13429.

Roessler, E., Ermilov, A. N., Grange, D. K., Wang, A., Grachtchouk, M., Dlugosz, A. A. & Muenke, M. 2005. A previously unidentified amino-terminal domain regulates transcriptional activity of wild-type and disease-associated human GLI2. *Hum Mol Genet,* **14,** 2181-2188.

Ruiz I Altaba, A. 1998. Combinatorial Gli gene function in floor plate and neuronal inductions by Sonic hedgehog. *Development*.

Ruppert Jm, K. K., Wong Aj, Bigner Sh, Kao Ft, Law Ml, Seuanez Hn, O'brien Sj, Vogelstein B. 1998. The GLI-Kruppel family of human genes. *Mol Cell Biol.***,** 3104-3113.

Sasaki H, N. Y., Hui C, Nakafuku M, Kondoh H. 1999. Regulation of Gli2 and Gli3 activities by an amino-terminal repression domain: implication of Gli2 and Gli3 as primary mediators of Shh signaling. *Development.,* **126,** 3915-3924.

Schimmang, T., Lemaistre, M., Vortkamp, A. & Ruther, U. 1992. Expression of the zinc finger gene Gli3 is affected in the morphogenetic mouse mutant extra-toes (Xt). *Development,* **116,** 799-804.

Schimmang, T., Oda, S. I. & Ruther, U. 1994. The mouse mutant Polydactyly Nagoya (Pdn) defines a novel allele of the zinc finger gene Gli3. *Mamm Genome,* **5,** 384-386.

Schweitzer, R., Vogan, K. J. & Tabin, C. J. 2000. Similar expression and regulation of Gli2 and Gli3 in the chick limb bud. *Mech Dev,* **98,** 171-174.

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jonsson, B., Schluter, D. & Kingsley, D. M. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature,* **428,** 717-723.

Shin, S. H., Kogerman, P., Lindstrom, E., Toftgard, R. & Biesecker, L. G. 1999. GLI3 mutations in human disorders mimic Drosophila cubitus interruptus protein functions and localization. *Proc Natl Acad Sci U S A,* **96,** 2880-2884.

Shubin, N., Tabin, C. & Carroll, S. 2009. Deep homology and the origins of evolutionary novelty. *Nature,* **457,** 818-823.

Shubin, N. H. & Dahn, R. D. 2004. Evolutionary biology: Lost and found. *Nature,* **428,** 703-704.

Sicklick, J. K., Li, Y. X., Jayaraman, A., Kannangai, R., Qi, Y., Vivekanandan, P., Ludlow, J. W., Owzar, K., Chen, W., Torbenson, M. S. *et al.* 2006. Dysregulation of the Hedgehog pathway in human hepatocarcinogenesis. *Carcinogenesis,* **27,** 748-757.

Spitz, F. 2001. Large scale transgenic and cluster deletion analysis of the HoxD complex separate an ancestral regulatory module from evolutionary innovations. *Genes & Development,* **15**.

Spitz, F., Gonzalez, F. & Duboule, D. 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell,* **113,** 405-417.

Stecca, B. & Altaba, A. R. I. 2010. Context-dependent Regulation of the GLI Code in Cancer by HEDGEHOG and Non-HEDGEHOG Signals. *Journal of Molecular Cell Biology,* **2,** 84-95.

Stern, D. & Orgogozo, V. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution; international journal of organic evolution,* **62,** 2155-2177.

Tanimoto, K., Liu, Q., Bungert, J. & Engel, J. D. 1999. Effects of altered gene order or orientation of the locus control region on human beta-globin gene expression in mice. *Nature,* **398,** 344-348.

Tanimura, A., Dan, S. & Yoshida, M. 1998. Cloning of novel isoforms of the human Gli2 oncogene and their activities to enhance tax-dependent transcription of the human T-cell leukemia virus type 1 genome. *J Virol,* **72,** 3958-3964.

Tarchini, B. & Duboule, D. 2006. Control of Hoxd genes' collinearity during early limb development. *Dev Cell,* **10,** 93-103.

Te Welscher, P., Zuniga, A., Kuijper, S., Drenth, T., Goedemans, H. J., Meijlink, F. & Zeller, R. 2002. Progression of vertebrate limb development through SHH-mediated counteraction of GLI3. *Science,* **298,** 827-830.

Theil, T., Kaesler, S., Grotewold, L., Bose, J. & Ruther, U. 1999. Gli genes and limb development. *Cell Tissue Res,* **296,** 75-83.

Thisse, C. & Thisse, B. 2005. High Throughput Expression Analysis of ZF-Models Consortium Clones. ZFIN Direct Data Submission (http://zfin.org).

Thomas, M. C. & Chiang, C. M. 2006. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol,* **41,** 105-178.

Thomas, N. A., Koudijs, M., Van Eeden, F. J., Joyner, A. L. & Yelon, D. 2008. Hedgehog signaling plays a cell-autonomous role in maximizing cardiac developmental potential. *Development,* **135,** 3789-3799.

Tojo, M., Kiyosawa, H., Iwatsuki, K., Nakamura, K. & Kaneko, F. 2003. Expression of the GLI2 oncogene and its isoforms in human basal cell carcinoma. *The British journal of dermatology,* **148,** 892-897.

Venkatesh, B., Tay, B. H., Elgar, G. & Brenner, S. 1996. Isolation, characterization and evolution of nine pufferfish (Fugu rubripes) actin genes. *J Mol Biol,* **259,** 655-665.

Vilar, J. M. & Saiz, L. 2005. DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise. *Curr Opin Genet Dev,* **15,** 136-144.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. *et al.* 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature,* **457,** 854-858.

Visel, A., Bristow, J. & Pennacchio, L. A. 2007. Enhancer identification through comparative genomics. *Seminars in Cell & Developmental Biology,* **18,** 140152.

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E. M. & Pennacchio, L. A. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet,* **40,** 158-160.

Voronova, A., Al Madhoun, A., Fischer, A., Shelton, M., Karamboulas, C. & Skerjanc, I. S. 2012. Gli2 and MEF2C activate each other's expression and function synergistically during cardiomyogenesis in vitro. *Nucleic Acids Res,* **40,** 3329-3347.

Vortkamp, A., Gessler, M. & Grzeschik, K. H. 1991. GLI3 zinc-finger gene interrupted by translocations in Greig syndrome families. *Nature,* **352,** 539-540.

Wang, X., Zhao, Z., Muller, J., Iyu, A., Khng, A. J., Guccione, E., Ruan, Y. & Ingham, P. W. 2013. Targeted inactivation and identification of targets of the Gli2a transcription factor in the zebrafish. *Biol Open,* **2,** 1203-1213.

Washington Smoak, I., Byrd, N. A., Abu-Issa, R., Goddeeris, M. M., Anderson, R., Morris, J., Yamamura, K., Klingensmith, J. & Meyers, E. N. 2005. Sonic hedgehog is required for cardiac outflow tract and neural crest cell development. *Dev Biol,* **283,** 357-372.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. & Young, R. A. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell,* **153,** 307-319.

Wittkopp, P. & Kalay, G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature reviews. Genetics,* **13,** 59-69.

Woolfe, A. & Elgar, G. 2007. Comparative genomics using Fugu reveals insights into regulatory subfunctionalization. *Genome biology,* **8**.

Woolfe, A. & Elgar, G. 2008. Organization of conserved elements near key developmental regulators in vertebrate genomes. *Advances in genetics,* **61,** 307-338.

Woolfe, A., Goode, D. K., Cooke, J., Callaway, H., Smith, S., Snell, P., Mcewen, G. K. & Elgar, G. 2007. CONDOR: a database resource of developmentally associated conserved non-coding elements. *BMC Dev Biol,* **7,** 100.

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., Mcewen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K. *et al.* 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol,* **3,** e7.

Yamada, T., Pfaff, S. L., Edlund, T. & Jessell, T. M. 1993. Control of cell pattern in the neural tube: motor neuron induction by diffusible factors from notochord and floor plate. *Cell,* **73,** 673-686.

Yanagisawa, H., Clouthier, D. E., Richardson, J. A., Charite, J. & Olson, E. N. 2003. Targeted deletion of a branchial arch-specific enhancer reveals a role of dHAND in craniofacial development. *Development,* **130,** 1069-1078.

Yue, S., Chen, Y. & Cheng, S. Y. 2009. Hedgehog signaling promotes the degradation of tumor suppressor Sufu through the ubiquitin-proteasome pathway. *Oncogene,* **28,** 492-499.

Zhang, J., Wagh, P., Guay, D., Sanchez-Pulido, L., Padhi, B. K., Korzh, V., Andrade-Navarro, M. A. & Akimenko, M. A. 2010. Loss of fish actinotrichia proteins and the fin-to-limb transition. *Nature,* **466,** 234-237.

APPENDIX

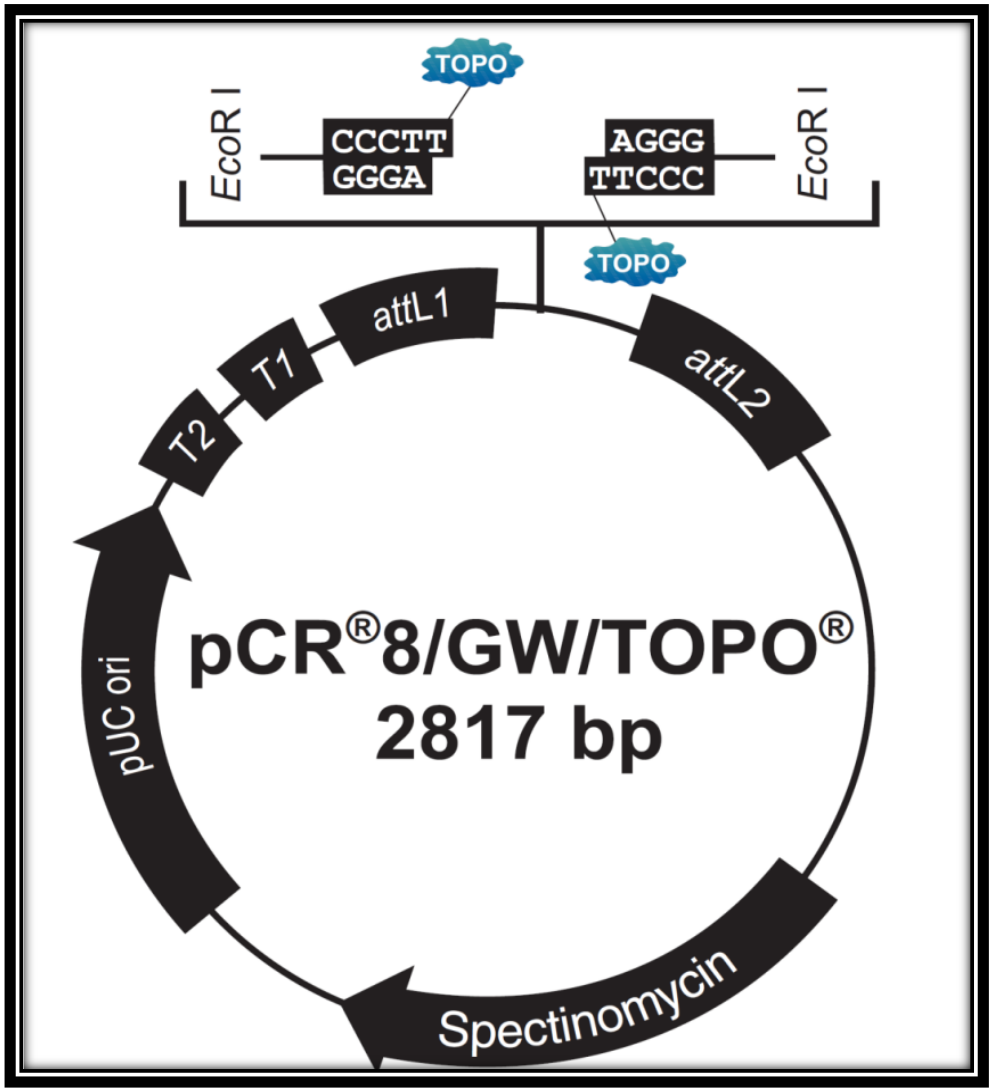

Figure A1. Map shows the features of the pCR™8/GW/TOPO® vector.

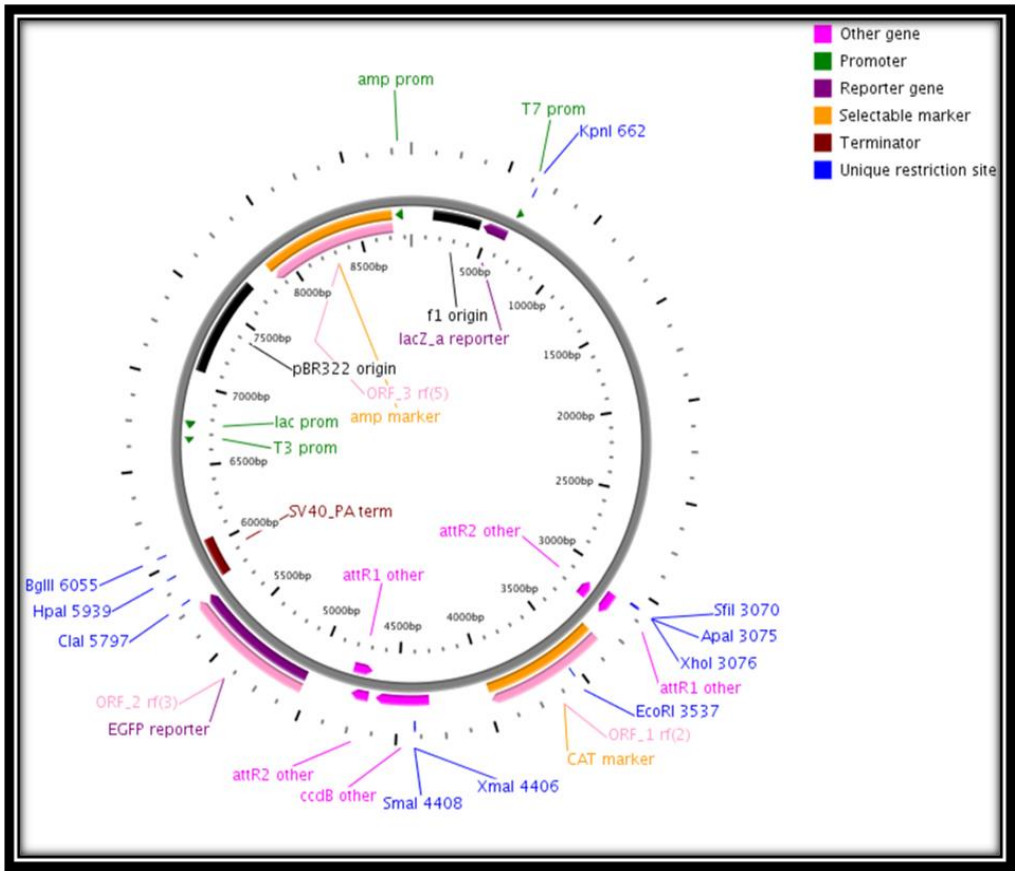

Figure A2. Map of the *cfos-IsceI*-EGFP plasmid.

Key features are highlighted and explained by the colour-coded key.

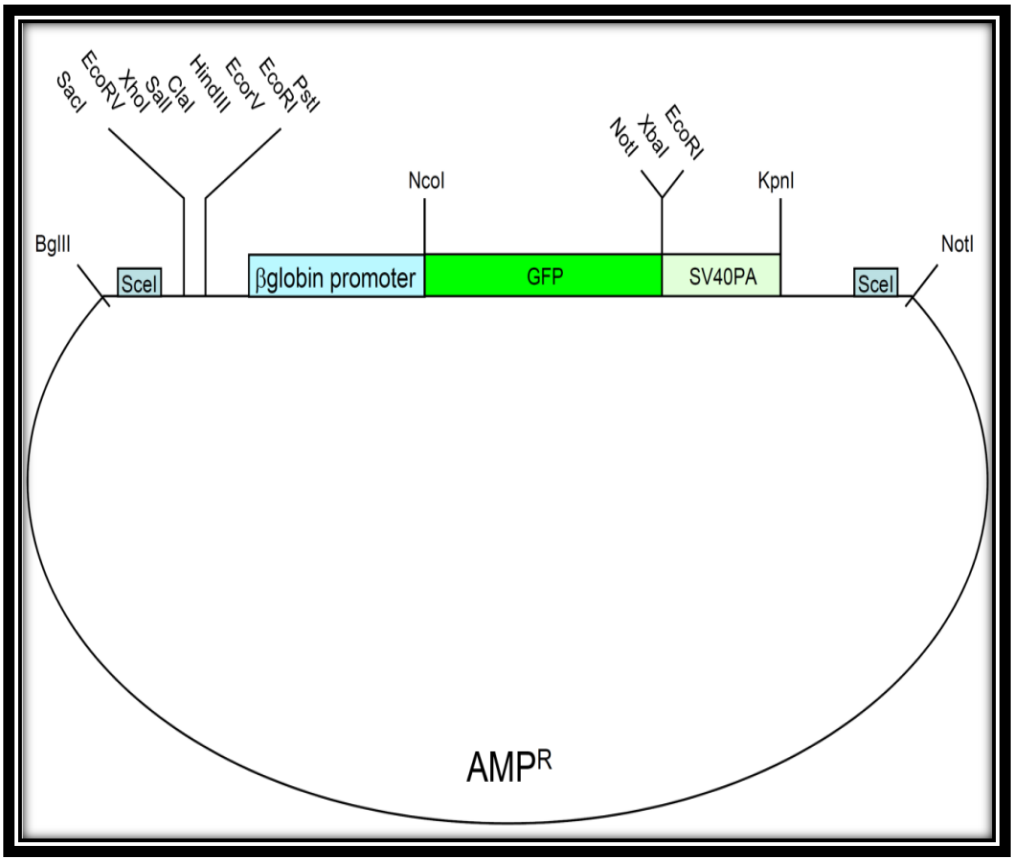

Figure A3. Map of the *β-globin* EGFP plasmid.

Multiple cloning sites and key features are highlighted.

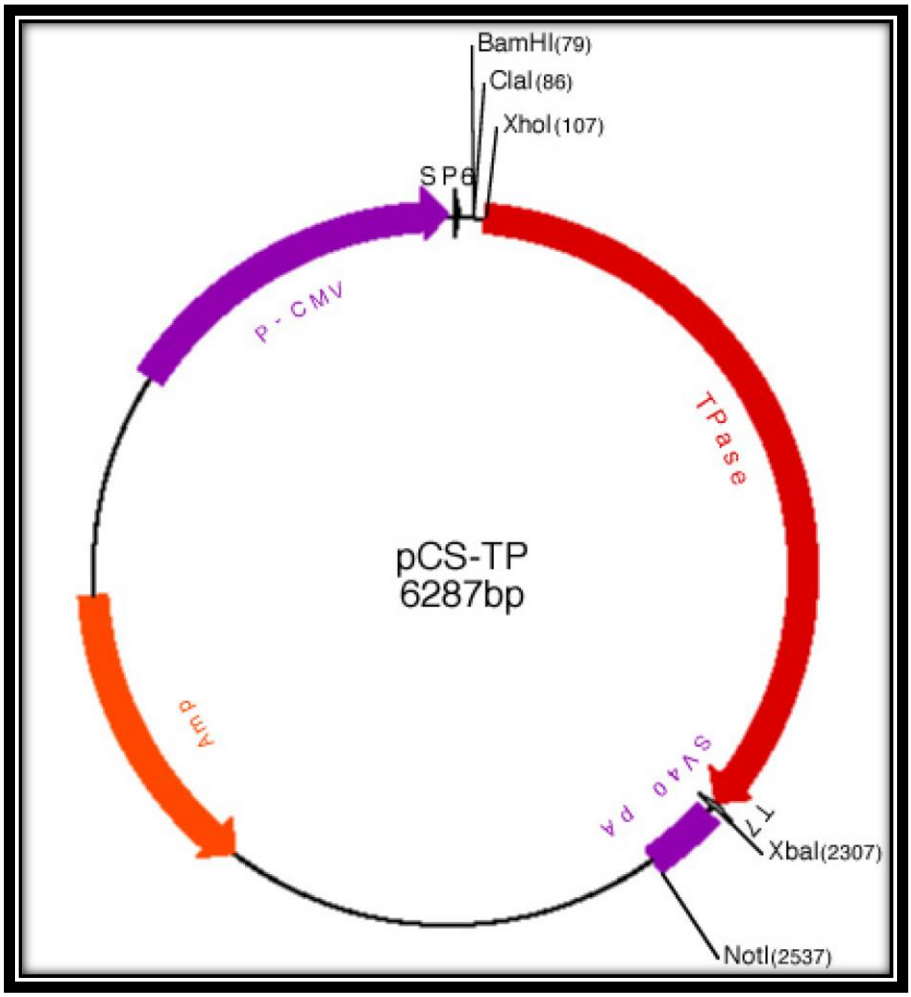

Figure A4. Map of the pCS-TP vector.

Key features are highlighted and explained. NotI restriction site was used to linearize the vector for *in vitro* transcription of RNA as described previously (Kawakami *et al.*, 2004).

Table A1 1 Plasmid-specific primers and their annealing temperatures

Element	Forward Primer	Reverse Primer	Annealing Temperature
β-globin EGFP	GGAAGGCCATCCAGCCTC	GTGCCACCTGACGTCTAAG	60˚C
Topo	GTTGCAACAAATTGATGAGCAATGC	GTTGCAACAAATTGATGAGCAATTA	58˚C
Tol2	TGTCTGAAACACAGGCCAGA	*	60˚C

***Tol2*-specific forward primer was used for the sequencing reaction of destination clone. No reverse primer was designed.**

PUBLICATIONS

1. Amina B., Anwar S., **Minhas, R**., Ali, S., Parveen, N., Azam, S.S., Nawaz, U. Abbasi, A.A.. Genomic features of human limb specific enhancers. (Manuscript submitted).
2. Anwar S., ***Minhas, R**., Ali, S., Lambertc, N., Kawakami, Y., Azam, S.S., Abbasi, A.A. 2015. Identification and characterization of novel transcriptional enhancers involved in regulating GLI3 expression during early development. **Dev Growth Differ 57(8): 570-580**. *** Co-first author**.
3. **Minhas, R**., Pauls, S., Ali, S., Doglio, L., Khan, M.K., Elgar, G., Abbasi, A.A. 2015. *Cis*-regulatory control of human GLI2 expression in the developing neural tube and limb bud. **Dev Dyn 244(5):681-92.**
4. Abbasi, A.A., **Minhas, R.**, Schmidt, A., Koch, S., Grzeschik, K.H. 2013. *Cis*-regulatory underpinnings of human GLI3 expression in embryonic craniofacial structures and internal organs. **Dev Growth Differ 55, 699-709.**
5. Parveen, N., Masood, A., Iftikhar, N., Minhas, B.F., **Minhas, R.,** Nawaz, U., Abbasi, A.A. 2013. Comparative genomics using teleost fish helps to systematically identify target gene bodies of functionally defined human enhancers. **BMC Genomics 14:122.**