GLI genes: *Cis*-Acting Regulatory Elements

Dissertation
zur
Erlangung des Doktorgrades
der Humanbiologie
(Dr. hum. biol.)

dem
Fachbereich Medizin
der Philipps-Universität Marburg, Germany

vorgelegt von

Amir Ali Abbasi

aus Birote, Pakistan

Zentrum für Humangenetik
Philipps-Universität
Marburg, 2008

GLI genes: *Cis*-Acting Regulatory Elements

Dissertation
zur
Erlangung des Doktorgrades
der Humanbiologie
(Dr. hum. biol.)

dem
Fachbereich Medizin
der Philipps-Universität Marburg, Germany

vorgelegt von

Amir Ali Abbasi

aus Birote, Pakistan

Zentrum für Humangenetik
Philipps-Universität
Marburg, 2008

Angenommen vom Fachbereich Humanmedizin
Der Philipps-Universität Marburg am.11-6-2008 (Tag der disputation)

Gedruckt mit Genehmigung des Fachbereichs

Dekan: Prof. Dr. med. M. Rothmund

Referent: Prof. Dr. rer . nat . K.-H. Grzeschik

Korreferent: Prof. Dr. Guntram Suske

CONTENTS

Zusammenfassung

Hintergrund: Frühembryonale Musterbildung wird durch zahlreiche komplexe Signalwege gesteuert. Funktionelle Interaktionen zwischen Komponenten einer bestimmten Signalübertragungs-Kaskade erfolgen auf vielen Ebenen, darunter bei der Bindung extrazellulärer Signalmoleküle an ihre Rezeptoren, beim Import der Signale von den Rezeptoren zum Zellkern oder bei der Interpretation eines Signals durch Aktivierung oder Repression der Expression von Zielgenen mittels Transkriptionsfaktoren. Präzise funktionelle Querverbindungen zwischen den Schritten einer Kaskade sind essentiell für eine normale Entwicklung.

Die Mitglieder der *GLI*-Genfamilie sind wichtige Vermittler der Signalinformation im "SONIC HEDGEHOG (SHH)-Signalweg". Während der letzten Jahrzehnte wurden die Mitglieder der GLI-Familie von Transkriptionsfaktoren eingehend mit genetischen, molekularen und biochemischen Methoden erforscht. Dadurch verfügen wir jetzt über reiche Informationen zur intrazellulären Lokalisierung der GLI-Proteine, über ihre Interaktionspartner, ihre Antwort auf SHH-Signale, über die Art ihres Transports in den Zellkern und auch über einige ihrer Zielproteine. Zudem wurde reiches Wissen über Entwicklungsstörungen beim Menschen und anderen Modellorganismen, die mit Mutationen von Mitgliedern der *GLI*-Genfamilie einhergehen, zusammengetragen.

In der vorliegenden Arbeit wurde die Bedeutung der GLI-Proteine für die Säugerentwicklung und die Evolution der Wirbeltiere untersucht, indem folgende Fragen beantwortet wurden:

I) Wie wird die Expression von *GLI*-Genen während der frühen embryonalen Entwicklung in Wirbeltierembryonen reguliert?

II) Nach welchen Mustern entwickelten sich die GLI-Proteine innerhalb einer Art und zwischen den Arten der Wirbeltiere.

III) Welche evolutionären Mechanismen haben dazu geführt, dass die paralogen *GLI*-Gene mit anderen Mitgliedern der *HOX*-Cluster-Paraloga in drei oder vier syntänen Abschnitten auf vier Chromosomen zusammen auftreten (beim Menschen auf den Chromosomen Hsa2, 7, 12 und 17)?

Ergebnisse und Folgerungen: Um genetische Mechanismen aufzuklären, mittels derer die Transkription von *GLI*-Genen während der frühen Embryonalentwicklung gesteuert wird, wurde in dieser Arbeit ein Familienmitglied, das menschliche *GLI*3-Gen, ausgewählt.

Durch evolutionären Sequenzvergleich zwischen zahlreichen Arten wurde in den Introns von *GLI*3 eine Architektur zwischen Tetrapoden und Teleostiern hochkonservierter, nicht-

kodierender Abschnitte entdeckt. Um darunter nach möglichen Enhancern der Genexpression zu fahnden, wurden 11 dieser konservierten nicht-kodierenden Elemente (CNEs) für eine funktionelle Analyse ausgewählt. Bei dieser wurde gezeigt, dass die hochkonservierten nicht-kodierenden Sequenzabschnitte in menschlichen Zelllinien und in der Embryogenese von Modellorganismen, Zebrafisch, Gliedmaßenentwicklung beim Hühnchen und in der Maus *GLI*3-spezifische regulatorische Aufgaben erfüllen konnten. Unter anderem hat diese Untersuchung zwei Enhancer definiert, die offenbar imstande sind, alle Aspekte der endogenen *GLI*3-Expression in knorpelbildenden und nicht-knorpelbildenden Mesenchymen embryonaler Gliedmaßen in einer nicht-redundanten Weise zu rekapitulieren. *In vivo*-Daten von Zebrafisch, Hühnchen und Maus deuten darauf hin, dass ein über lange Zeiträume hin konservierter Enhancer im Verlauf der Evolution dafür umfunktioniert wurde, die *GLI*3-Expression in "moderneren" Abschnitten der Gliedmaßenstruktur, wie in Händen und Füßen (Autopodia), zu steuern. Demgegenüber behielt eine zweite Gliedmaßen-spezifische Enhancerregion ihre ursprünglichen Funktionen bei, die Expression von *GLI*3 in "alten" Bereichen von Flossen/Gliedmaßen, dem Stylopod und dem Zeugopod festzulegen.
Um für kodierende Bereiche der *GLI*-Gene Muster der evolutionären Entwicklung aufzuschlüsseln, wurde eine molekulare Analyse *in silico* durchgeführt, in die GLI-Sequenzen representativer Mitglieder der Tetrapoden- und Teleostier-Entwicklungslinien einbezogen wurden. Diese Untersuchung bestätigte, dass die Veränderungen, denen die GLI-Sequenzen im Verlauf der Evolution unterlagen, grundsätzlich mit den bekannten funktionellen Ähnlichkeiten und Unterschieden innerhalb und zwischen den Arten zusammenpassen.

Weiterhin wurde in der vorliegenden Arbeit der Versuch unternommen, die Vorgänge im Verlauf der Evolution zu beleuchten, die die menschlichen paralogen *GLI*-Gene und Mitglieder mehrerer anderer Genfamilien in der Nähe der *HOX*-Cluster auf vier Chromosomen jeweils in drei oder vier kollinearen Abschnitten zusammengeführt haben (beim Menschen auf den Chromosomen 2, 7, 12 und 17). Dazu wurde die phylogenetische Geschichte von 11 Multigenfamilien analysiert, von denen drei oder mehr Mitglieder mit den menschlichen *HOX*-Clustern gekoppelt sind. Die Ergebnisse dieser Untersuchung deuten darauf hin, dass die heute beobachtete umfangreiche Syntänie auf drei oder vier menschlichen *HOX*-Cluster-tragenden Chromosomen das Ergebnis früherer kleiner Duplikationen (von Segmenten oder Gengruppen) widerspiegelt, auf die zu unterschiedlichen Zeiten während der Chordatenevolution genomische Rearrangements folgten.

Signifikanz: Passend zur komplexen Rolle von GLI3 bei einer Fülle von Schritten der Musterbildung während der Embryonalentwicklung der Vertebraten hat die vorliegende Untersuchung ein altes regulatorisches Netzwerk aus mehreren spezifisch wirkenden cis-agierenden Elementen entdeckt. Die Beschreibung eines Katalogs *GLI*3-spezifischer *cis*-regulatorischer Elemente eröffnet insbesondere eine neue Perspektive, die genetischen

Mechanismen zu verstehen, über die Effektoren der SHH-Signalkaskade selbst zur rechten Zeit am richtigen Ort bereitgestellt werden, um dann während der Embryogenese die Musterbildung entlang der Körperachsen zu steuern. Die Ergebnisse zeigen zum Beispiel einen Mechanismus auf, über den in benachbarten Domänen der Gliedmaßenknospen, des sich entwickelnden Neuralrohrs oder in der Organogenese die richtige Balance zwischen den Transkripten von *SHH* und *GLI*3 etabliert wird. Zudem könnten diese *cis*-regulatorischen Elemente zur Erklärung möglicherweise *GLI*3-assoziierter Entwicklungsstörungen bei Patienten herangezogen werden, bei denen keine Mutation in kodierenden Abschnitten des *GLI*3-Gens nachweisbar ist. In diesen Fällen bietet es sich jetzt an, diese Enhancer auf Mutationen durchzusuchen, die die Verfügbarkeit von *GLI*3-Transkripten während der Embryogenese beeinträchtigen.

Die genetischen Mechanismen, mit deren Hilfe sich die Körperanhangsstrukturen der Wirbeltiere herausgebildet und schrittweise zu Vorder- und Hinterbeinen der modernen terrestrischen Vertebraten entwickelt haben, liegen noch weithin im Dunkeln. Die vorliegende Untersuchung erweitert den Wissensstand darüber, wie sich die moderneren Bereiche der Gliedmaßen, wie zum Beispiel die Autopodia entwickelt haben könnten. Die Eingrenzung zweier unterschiedlicher Enhancer, die unabhängig von einander die *GLI*3-Expression in den proximalen und distalen Bereichen der Gliedmaßen steuern, deutet an, dass in neuerer Zeit der Vertebratenevolution GLI3- Funktionen durch Umwidmung eines alten Enhancerelements dazu benutzt wurden, die Entwicklung von Anzahl und Identität der Finger und Zehen zu kontrollieren.

Die Analyse der molekularen Evolution *GLI*-kodierender Sequenzen hat allgemeine Bedeutung in dem Sinne, dass die hier erzielten Ergebnisse frühere Befunde aus der Untersuchung anderer Vertebratengene stützen, nach denen sich Gene in Fischen schneller weiter entwickeln als in Säugetieren, und dass Paraloge sich in typischer Weise mit gleicher Geschwindigkeit entwickeln, ohne signifikante Asymmetrie. Weiterhin zeigt die Übereinstimmung der Befunde aus der evolutionären Analyse des Sequenzmusters der *GLI*-Familienmitglieder mit denen, die aus funktionellen Untersuchungen dieser Gene bei verschiedenen Vertebratenarten stammen, dass die molekulare Analyse der Evolution von Genfamilien in silico: i) eine Vorausschau auf mögliche funktionelle Verwandtschaft zwischen Genen erlaubt, ii) als wertvolle Handreichung zur Planung weiterer experimenteller Ansätze zur funktionellen Charakterisierung von Mitgliedern einer Genfamilie in unterschiedlichen Arten dienen kann oder iii) die Ergebnisse funktioneller Studien absichern kann.

Es wurde angenommen, dass die vierfach vorhandenen Paralogieregionen (Paraloga) im menschlichen Genom, insbesondere auf den Chromosomen HSA1/6/9/19, HSA4/5/8/10 und auf den *HOX*-tragenden Chromosomen HSA2/7/12/17 direkt durch zwei

Polyploidisierungsrunden entstanden sind (2R-Hypothese). Die Schlussfolgerung, dass die Koppelungsbeziehungen, die man auf den *HOX*-tragenden Chromosomen beobachtet, nicht auf zwei Polyploidisierungsrunden zurückzuführen sind, kann bedeutende Auswirkungen auf die Klärung der kontroversen Ansichten über die evolutionären Prozesse haben, die unser Genom geformt haben.

Insgesamt weisen die in dieser Dissertation dargelegten Ergebnisse für das *GLI*3-Gen einen äußerst komplexen,evolutionär hochkonservierten *cis*-aktiven Regulationsmechanismus nach, und sie zeigen, wie dieser Bestand an *cis*-regulatorischen Elementen herangezogen werden konnte, um die Komplexität des Körperbauplans durch Umwidmung von *GLI*3-Funktionen für neue Entwicklungsaufgaben zu erweitern. Darüber hinaus bietet diese Arbeit auch neue Einsichten zu evolutionären Aspekten der codierenden *GLI*-Sequenzen und deren Bedeutung für die Evolution unseres Genoms.

Summary

Background: Early embryonic patterning is regulated by many complex signaling pathways. The functional connections among components of a particular signal transduction pathway occur at many levels, including the interaction of extra-cellular signaling molecules with their receptors, the import of signals from receptors to the nucleus, the interpretation of a signal by the activation or repression of target gene expression through transcription factors. The precise functional connectedness among members of a cascade is essential for normal development. The members of the *GLI* gene family of transcription factors are key mediators of one such pathway known as "SONIC HEDGEHOG (SHH)" signaling pathway. During the past couple of decades GLI family members have been extensively scrutinized by genetic, molecular and biochemical means. Thus, a wealth of information is currently available about the intracellular localization of GLI proteins, their interacting partners, their response to SHH signals, the way they are transported to nucleus, and about some of their downstream target genes. In addition, a great deal of information about the developmental defects associated with mutations in *GLI* gene family members in humans and other model organisms has emerged.

In this study, the importance of GLI proteins in mammalian development and vertebrate evolution has been explored by answering the following questions:

I) How is the expression of *GLI* genes regulated during early developmental patterning of vertebrate embryos?

II) What are the patterns, along which the *GLI* proteins evolved within and between vertebrate species?

III) What are the evolutionary mechanisms, that combined the *GLI* paralogs and other members of the *HOX* cluster paralogon in three or four collinear regions on different chromosomes (Hsa2, 7, 12, and 17)?

Results & Conclusions: Towards the elucidation of genetic mechanisms, by which the transcription of *GLI* genes is regulated during early embryonic development, in this study one of the members of this family, i.e. human *GLI3,* was selected.

By employing multispecies sequence alignment, an anciently conserved (tetrapod-teleost) non-coding architecture within the introns of *GLI3* was identified. To search for possible enhancers of expression, 11 of these tetrapod-teleost conserved non-coding elements (CNEs) were selected for functional analysis. A *GLI3* specific regulatory function of these deeply conserved intronic sequences was detected in human cell lines and model organisms: zebrafish, chicken (limb-bud), and mouse.

In particular, this study has defined two distinct enhancers apparently recapitulating the entire known aspects of endogenous *GLI*3 expression within cartilaginous and non cartilaginous mesenchyme of embryonic limbs in a non-redundant manner. *In vivo* data from zebrafish, chicken and mice suggests that one limb specific anciently conserved enhancer might have been co-opted during the course of evolution to regulate *GLI*3 expression within more modern aspects of vertebrate appendicular structure, i.e. within hands and feet (autopodia). In contrast with respect to fin/limb specificity a second limb specific enhancer region preserved its ancient functions, dictating *GLI*3 expression within ancient fin/limb domains, i.e. stylopod and zeugopod.

To sort out the evolutionary patterns of *GLI* coding sequences, a molecular evolutionary analysis *in silico* was carried out employing GLI sequences from representative members of tetrapod and teleost lineages. This analysis confirmed that changes experienced by the GLI sequences during the course of vertebrate evolution are largely in agreement with the reported similarities and differences in their functions within and between the species.

In the present study, in an attempt to elucidate those ancient evolutionary events which brought the human *GLI* paralogs and members of many other gene families in the physical proximity of *HOX* clusters in three or four collinear regions of different chromosomes (Hsa2, 7, 12 and 17; *HOX* cluster paralogon), the phylogenetic history of 11 multigene families with three or more of their representatives linked to human *HOX* clusters was analyzed. The results from this analysis suggest that that extensive triplicate or quadruplicate synteny that is seen on the present day human *HOX*-bearing chromosomes is the result of ancient small-scale duplications (segmental or gene-clusters) and subsequent genomic rearrangement events which occurred at different time points during chordate evolution.

Significance: Congruent with the complex role of GLI3 in a multitude of patterning steps during vertebrate embryonic development, in this study, an ancient regulatory network comprising multiple distinctly acting cis-acting elements was revealed.

In particular, the elucidation of a *GLI3* specific *cis*-regulatory catalog offers a new perspective on understanding the genetic mechanisms by which the downstream effectors of SHH signaling cascade might themselves be regulated at correct place and precise time to direct pattern formation along the body axis during embryogenesis. For instance, this data could help to understand the mechanisms by which a proper balance between *SHH* and *GLI*3 transcripts is established in complementary domains within the developing limb and neural tube, and also during organogenesis. In addition, these *cis*-regulatory elements might contribute to understanding the genetic basis of those potentially *GLI3*-associated human birth defects which cannot be attributed to a mutation in the coding sequence of this gene. In such cases, these enhancers can be searched for mutations that can potentially affect the availability of *GLI3* transcripts during embryogenesis.

The genetic mechanisms by which the vertebrate appendicular structures emerged and subsequently evolved into forelimbs and hind limbs in modern terrestrial vertebrates remain largely elusive. This study contributes to our current knowledge of how the more modern aspects of limbs such as autopodia might have evolved. The localizing of two distinct enhancers controlling *GLI3* expression along the proximal and distal extremities of limbs, independently, suggests that *GLI*3 functions might have been recruited to impose developmental constraints on digit number and identity late in vertebrate evolution through the co-option of an ancient enhancer element.

The molecular evolutionary analysis of *GLI* coding sequences has general implications in a sense, that results acquired in this study provide support to the previous data obtained from the analysis of other vertebrate genes, i.e. that genes evolve faster in fish than in mammals and that paralogs typically evolve at similar rates, without significant asymmetry. In addition, the harmony between the data obtained from the analysis of sequence evolutionary patterns of *GLI* family members and those that are generated through functional investigations of these genes in various vertebrate species suggests, that molecular evolutionary analysis of gene families through computational means, i) can offer a glimpse at the putative functional relatedness among genes, ii) can serve as a valuable guide to design future experimental approaches for the functional characterizations of gene family members in different species, or iii) can validate the results obtained through functional approaches.

The four-fold paralogy regions (paralogons) in the human genome, notably on HSA 1/6/9/19, HSA 4/5/8/10, HSA 1/2/8/10, and the *HOX*-bearing chromosomes HSA 2/7/12/17 are considered to be shaped directly by two rounds of polyploidization. The conclusion that the linkage relationships seen on the human *HOX*-bearing chromosomes are not an outcome of two rounds of polyploidization events (2R hypothesis) may have important implications in resolving the controversies about the evolutionary processes that had shaped our own genome.

Taken together, the data presented in this thesis reveal a highly complex and evolutionarily deeply conserved cis-regulatory control of human *GLI3* gene and shows how this cis-acting catalogue might have contributed towards the complexity in vertebrate body plan through recruitment of GLI3 functions for novel developmental tasks. In addition, this study also provides an insight into the evolutionary aspects of *GLI* coding sequences and their relevance to the evolution of our genome.

INTRODUCTION

The early events in vertebrate development generate modularity in the body plan by dissociating the embryo into developmentally autonomous compartments. This modularity in turn allowed the local, developmentally autonomous embryonic structures to accommodate genetic and molecular variations without altering the developmental events within adjoining compartments. This regionally restricted accumulation of variations without affecting the general body plan conferred on animals the ability to experience morphological innovations leading to great morphological diversity.

Genetically determined morphological diversity could be based on amplification and mutation of coding genetic material expanding the genetic toolkit for development and/or the expansion and evolution of regulatory components that act *in cis* to control expression of developmental regulators. Within well studied animal subgroups, such as the mammals among the vertebrates, the variability in the overall contents of nuclear genetic material is minimal implicating a crucial role for evolution of regulatory elements.

With an increasing body of empirical evidence emerging from the fields of *evolutionary developmental biology* and *comparative genomics*, it is now broadly accepted that increased morphological complexity and diversity in mammals is associated with the evolution of *cis*-acting DNA elements that regulate the expression of developmental regulators.

Unlike the coding sequences where insertions, deletions, or substitutions within non-synonymous sites are often intolerable, the *cis*-acting regulatory modules exhibit plasticity. They often harbor short (4-6 bp) degenerate binding sites for multiple *trans*-acting factors. Combinatorial, simple base pair changes within *cis*-acting DNA can potentially alter the binding affinities for existing factors, while insertions and deletions can alter the site spacing, create new or delete existing binding sites. Thus contrary to coding sequences, many sequence variations within *cis*-acting regulatory elements potentially impose tolerable affects on their activity resulting in incremental variations in timing, pattern and level of the expression of the associated gene, and can work as a fuel for evolution in morphological diversity and complexity in several ways.

I) Innovation of an enhancer directing the deployment of the associated developmental regulator to a domain where it was not previously expressed. This expansion in *cis*-acting regulatory contents can potentially expand the functional territories of the associated coding regions and thus create increasingly complex developmental compartments through pleiotropy in the usage of the existing genetic toolkit. *Cis*-acting regulatory element-triggered

pleiotropy of developmental gene expression eliminates the need for expansion in number and types of developmental regulators for morphological and anatomical evolution.

II) Novel specificity of *cis*-acting regulatory sequences. Developmental genes are associated with an array of transcriptional enhancers each of which can potentially dictate their expression in a subset of tissues. Sequence changes of individual transcription factor binding sites within enhancer elements or alterations in the molecular anatomy of enhancers will have regionally localized effects on the phenotype which may not have deleterious consequences on the over all fitness.

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 development of paired appendages.

In vertebrates, these are serially iterated structures (fins in fish and limbs in tetrapods), and during evolution there has been a trend towards their morphological and functional diversification both within and between taxa. The vertebrate appendages have been in the focus of intense genetic and molecular investigations, because they are not essential for embryonic survival and can be experimentally manipulated in model organisms to define the important cellular and molecular interactions that regulate patterning and skeletal development. The elucidation of many crucial genes that regulate growth and patterning of limbs has now made it evident, that despite of their extreme morphological and functional diversification from fish fin to human limb, the vertebrate appendicular architecture is build upon a fairly similar repertoire of regulatory genes. A growing body of evidence from detailed studies on a subset of limb regulators, like the *HOXD* cluster or *SHH*, suggests that evolution of the regulatory components was the key for the origin and subsequent morphological diversification of the vertebrate appendicular skeleton. However, the picture is still incomplete as we know little about the *cis*-acting regulatory contents of other crucial genes involved in limb patterning and development. The detection and functional analysis of *cis*-acting regulatory elements of key developmental regulators like *GLI3* is thus an essential contribution towards a composite map of evolutionary events involved in the amazing architectural and functional diversification of limbs between tetrapod and fish lineages and within tetrapods.

1.1 Evolution from simple to complex organisms

The size and shape of life did not expand appreciably until the invention of bilaterians in late Proterozoic era comprising most animals from worms to humans. The characteristics feature that differentiates bilaterians from other forms of life is their complex body

organization with well defined rostral-caudal and dorsal-ventral axis and a plane of mirror symmetry running between left and right sides (bilateral symmetry). In contrast, the simple metazoans like Porifera (sponges) and Cnidaria (jelly fish and sea anemones) possess 10-12 cell types and exhibit a simple form of symmetry termed radial symmetry. These simple forms of life are diploblastic and diverged early in animal evolution before the invention of triploblastic bilaterians, with a third, mesodermal germ layer. The evolution of mesoderm and its derivatives had profound consequences on organismal complexity in terms of overall size and locomotion. Before the Cambrian era, the bilaterally symmetrical triploblastic life forms splitted into two major lineages, protostomes and deuterostomes. The major distinctions between these two lineages are found in embryonic development. In protostomes the oral end (mouth) of the animal develops from the first developmental opening, the blastopore. The majority of the coelomate invertebrates develop as protostomes. In deuterostomes including Echinoderms and the ancestors of the Chordates, the oral end of the animal develops from a second opening on the dorsal surface of the animal, and the blastopore becomes the anus. Furthermore, the deuterostomes show indeterminate development. In contrast, the protostomes undergo determinate development in which the developmental fate of each cell in the adult organism has already been defined.

During the course of ~2000 millions of years of parallel evolutionary time (molecular clock time estimates) both protostomes (such as the arthropod, *Drosophila melanogaster*) and deuterostomes (e.g. vertebrates) have attained modular body plans, great morphological complexity, and taxonomic diversity. One particular type of complexity of great interest is that of occurrence of serially repeated structures. Body segments in arthropods, vertebrae in vertebrates, limbs in both taxa, and teeth are serially homologous structures. The importance of the construction of the animal body plan from repeated parts has long been recognized (McShea 1991), and serially repeated modular construction is considered to have various advantages including the facilitation of great size and efficiency as well as the evolution of greater complexity and adaptation through the functional differentiation of repeated parts (Carroll 2001).

In fact, the innovation of serially homologous paired appendages in vertebrates and their morphological specialization into forelimbs and hind limbs in demands of feeding and locomotion in late Devonian (400-350Mya) led to bursts of diversification of tetrapod lineage, distributed almost all over the terrestrial ecosystem.

Similarly in the evolution of diverse arthropods from lobopodan ancestors (Budd 1996), the mean and maximum number of serially homologous distinct limb-pair types increased

(Cisne 1974) with morphological specializations for feeding, locomotion, sensation, copulation, brooding young, burrowing, and defense.

The role of serially homologous appendages in adaptation and taxonomic diversification can be explained by the statement "The most specialized orders, those with the greatest number of different limb types, are also the most diverse in terms of the number of species" (Carroll 2001).

1.2 Mechanisms to generate morphological complexity

The expansion in the genetic toolkit and its direct consequence on an over all increase in morphological diversity is fit well at the base of deuterostomes. For instance there had been a trend in increase in number of coding contents during transition from their amphioxus like invertebrate ancestor to vertebrate lineage. This expansion in turn might have worked as a catalyst to provoke a burst in morphological diversity and complexity at the base of vertebrates, perhaps through tolerance of genetic changes in an initially redundant set of paralogs.

The gradual increase in overall morphological complexity and the innovation of novel structures like limbs during the evolutionary history of metazoa cannot be attributed entirely to the innovation of new gene contents. For example, the greater number of genes in *C.elegans* (~20.000) compared to *D.melanogaster* (~14.000) and in *Zebrafish* (~30.000) compared to humans (~25.000) does not reflect the index of morphological complexity (indeed *Drosophila* is more complex then *C.elegans,* and humans are more complex than *Zebrafish*). Since the time they originated from an invertebrate like common ancestor (~700 millions of years ago, *Mya*), the different vertebrate lineages from fish to human share a more or less similar catalog of developmental regulator genes. However, despite the extensive similarities in over all coding contents of their genomes, the vertebrates attained a surprising spectrum of morphological diversity and complexity.

The increase in morphological complexity has been attributed to the increase in number of transcription factors (Levine and Tjian 2003) and also to the expansion in number and complexity of regulatory elements that act *in cis* to control gene expression (Carroll 2001). The unicellular yeast genome encodes a total of ~300 transcription factors, while the genome sequences of *C.elegans* and *Drosophila* reveal at least 1000 transcription factors, each (Aoyagi and Wassarman 2000; Ruvkun and Hobert 1998). There may be ~3000 transcription factors in humans (Lander et al. 2001). Similarly, morphologically simple animals like *Ciona intestinalis* have been estimated to contain 10.000-20.000 tissue specific enhancers (Harafuji et al. 2002). *Drosophila* contains several enhancers per gene scattered over an average

distance of 10 kb. It is difficult to estimate the *cis*-regulatory contents of the human genome, however 5% have been estimated to be evolving more slowly than the neutral rate (Waterston et al. 2002) and less than 2% actually encode proteins. The remaining non-coding, slowly evolving sequences (~90 Mb) of the human genome are predicted to regulate temporal, spatial, and quantitative aspects of gene expression (Poulin et al. 2005; Waterston et al. 2002), among other roles.

The regulatory evolution in terms of increase in number of transcription factors and expansion in quantity and architectural complexity of cis-acting regulatory elements created new combinations of gene expression. To unveil these mechanisms is therefore key to understanding how the overall morphological complexity of animals has been achieved, in particular, how novel structures like limbs have evolved in the history of bilaterian evolution.

1.3 How to identify regulatory sequences?

The annotation of large genomes such as the human chromosomes for non-coding functional sequences, especially the gene regulatory landscapes, has been hampered by the fact that, unlike coding regions which have well characterized structural codes that allowed the computational scientists to design sophisticated algorithms to capture them with confidence, the *cis*-acting regulatory sequences have no definite vocabulary to which we are familiar to date. Extracting regulatory regions from the bulk of human genome (95%, ~2855 Mb) has become a major challenge for bioinformaticians and biologists.

Traditional ways of hunting for the gene regulatory elements involve various trial and error based methodologies. These classical experimental strategies were mostly confined to minimal promoter regions and promoter-proximal elements and include random cloning of target regions and subsequent deletion mapping to refine critical modules through a cell based reporter assay. DNAse1 hypersensitivity/EMSA (electrophoretic mobility shift assay) studies identify sequences that can potentially interact with transcription factors. *Cis*-acting regions have also been identified *in vivo* through BAC targeted enhancer trap approaches in transgenic mice. Most of the classical approaches are unguided and thus extremely laborious and time consuming.

The availability of the genome sequence of a variety of vertebrate organisms provided a powerful way to elucidate the gene regulatory networks through cross-species sequence comparisons. The rationale behind this approach is that functional sequences are subject to evolutionary constraints, which can leave conservation footprints in the aligned orthologous intervals, while neutrally evolving sequences change more rapidly.

Comparison of the human genome with phylogenetically close genomes like eutherian mammals (e.g. rodents, carnivores, and artiodactyls) or distantly related species like noneutherian mammals (marsupials and momotremes), birds, amphibians, and fish, each, has different advantages as well as shortcomings (Fig. 1.1). For example, the human and mouse pairwise comparison is found to detect an enormous degree of non-coding sequence similarities, probably due to an uneven rate of evolution across the genomes and lack of sufficient divergence time between eutherian mammals (Fig. 1.1). This background conservation makes it difficult to discern neutrally evolving non-coding sequences from functionally constrained ones. Despite of these shortcomings the human/mouse sequence comparisons aided in identification of numbers of conserved cis-acting regulatory elements (Gottgens et al. 2000; Loots and Ovcharenko 2004; Poulin et al. 2005; Waterston et al. 2002).

The distantly related Japanese pufferfish (*Fugu rubripes*) turned out to be an excellent model organism to annotate the human genome for *cis*-acting regulatory intervals because of

i) its small genome size (~400 Mb) due to reduced intergenic and intronic intervals,
ii) a similar gene repertoire to mammals,
iii) its extreme phylogenetic separation from mammals (~450 Mya) which assumably was sufficient to leave only functional regions to be conserved in mammalian-fish sequence comparison.

In fact, comparisons with the compact *Fugu* genome have identified a great number of anciently conserved regulatory sequences in mammals (Aparicio et al. 1995; Lettice et al. 2003; Nobrega et al. 2003). However, it should be noted that the sequence comparison at extreme phylogenetic separation (human/fish) has limitations in a sense, that such alignments will expose only the subset of regulatory components that is fundamentally significant for vertebrate embryogenesis and physiology (Fig. 1.1). Conversely, regulatory components for anatomical details that are specific to the mammalian lineage will have evolved subsequent to the divergence of fish and might be missed in mammals-fish comparison.

In order to overcome the shortcomings of pairwise comparison at moderate and extreme phylogenetic distance, an intermediate approach is advisable, i.e. the use of multiple orthologous sequences (>20 species), simultaneously, in a single comparison because

i) representing a range of organisms with moderate (eutherian mammals) and large branch lengths (marsupials, monotremes, birds, amphibians and fish) can provide a broad phylogenetic window to capture vertebrate specific, tetrapod specific, and mammalian specific, human conserved regulatory elements in a single comparison,
ii) increasing the numbers of species makes it progressively less probable that sequences are conserved by chance and helps to prioritize truly functional regulatory sequences

for experimental analysis,

iii) multi-species sequence comparison increases the alignability and can capture more conserved orthologous sequences than corresponding pairwise alignment, probably because increasing in diversity of sequences enhances statistical power for aligning those sequences that are too diverged to align pairwise (Brudno et al. 2003b; Margulies et al. 2006).

Currently, the public availability of genomic data from large numbers of vertebrate species (Fig. 1.1) separated from humans at phylogenetic distance of choice provides an unprecedented opportunity to make use of the immense potential of multi-sequence alignments for prioritizing human conserved elements for functional analysis to test their gene regulatory potential.

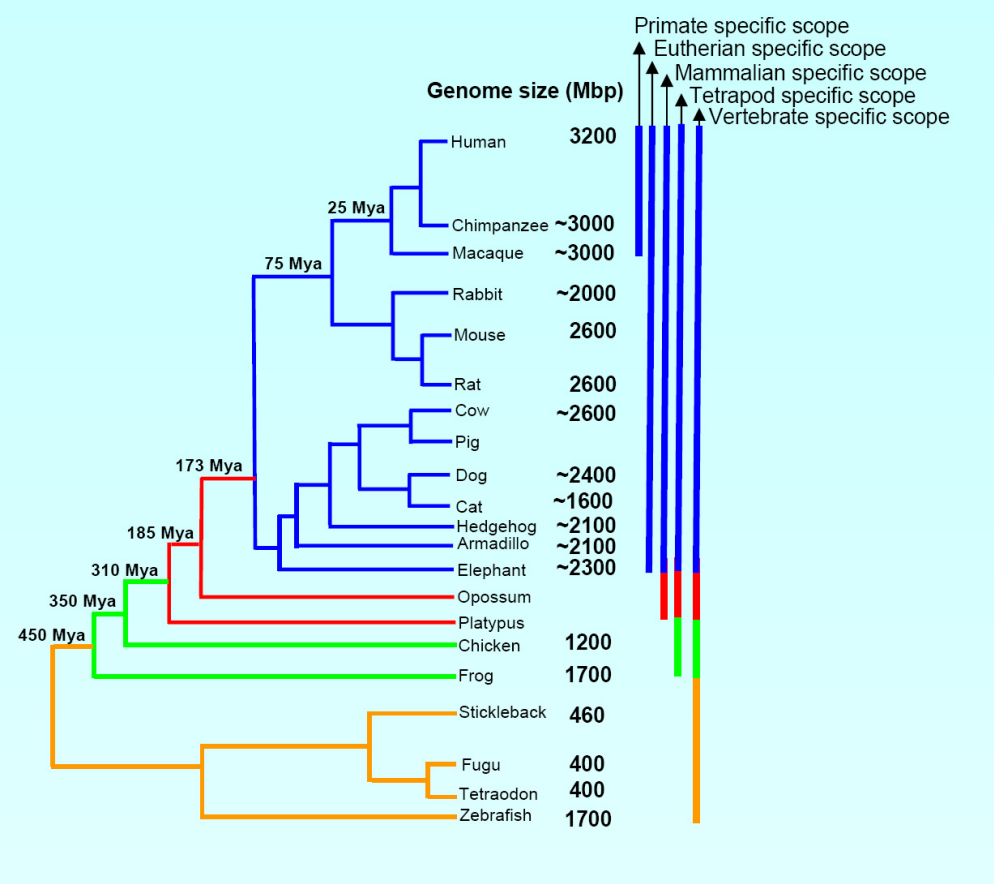


Figure. 1.1. Sequencing and the public availability of increasing number of vertebrate genomes provides an unprecedented opportunity to discern previously unknown functional landscapes in human genome through comparative sequence analysis.

Diagram illustrating the phylogenetic relationship of different vertebrate organisms whose genomes have been sequenced or are currently undergoing systematic sequencing. Approximate genome size is given in millions of base pairs (Mbp). Within the phylogenetic tree, blue, red, green, and golden color bars depict eutherian mammals, non-eutherian mammals, non-mammalian tetrapod, and fish lineages, respectively. On the right side of the diagram, the colored vertical bars represent the scopes of comparative sequence analysis at various phylogenetic separations to elucidate human conserved gene regulatory elements.

1.4 Sequence alignment tools

Sequence alignment is an arrangement of two or more sequences, to highlight their similarities (Fig. 1.2). Preferably, an alignment procedure should reflect the evolutionary relationship between two or more homologous sequences (evolved from common ancestor) and hint at those events that each of the homologous sequences had undergone independently since divergence from last common ancestor (Fig. 1.2).

Two major kinds of changes can occur at any given position within a sequence,

i) simple edits: that is insertion, deletion and substitutions of individual base pairs,

ii) rearrangement events: such as translocations (a subsegment is removed and positioned within a different location), inversion: (the orientation of a subsegment undergoes inversion while keeping its spatial position intact), duplication (original subsegment remains unchanged, while its copy is inserted).

In addition to these potential changes, homologous sequences can share certain subsegments of their sequences at base pair level and sometime in order and orientation because:

i) evolutionary constraints on functionally critical ancestral subsegments can make them immune to changes in descendants, even over the course of million of years,

ii) or the homologs depict similarities at neutral sites (functionally irrelevant) because of lack of sufficient divergence time (one of the ancestral black boxes in Fig. 1.2).

There are two broad sequence alignment strategies to capture the similarities between evolutionary related sequences (evolved from common ancestor):

i) global alignments; are produced when the two sequences are compared over their full length to determine the optimal similarity scores,

ii) local alignments; are generated when only related intervals of two sequences are compared and an optimal similarity score is determined over numerous alignable subregions along the length of two sequences.

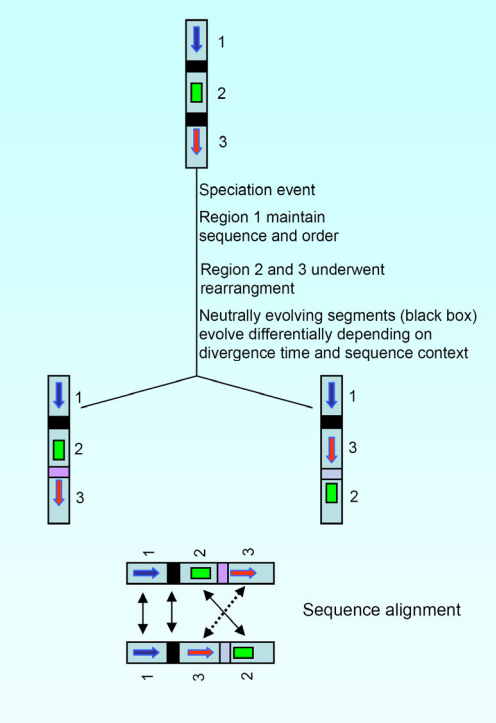


Figure. 1.2. Sequence alignments, if performed carefully lead not only to the elucidation of functional intervals, but also illuminate ancient evolutionary events.

Diagram illustrating the speciation event by which an ancestral sequence is passed on to two different lineages (bottom). The homologous intervals accumulate changes differentially in each lineage, with functionally critical subsegments (blue/red arrows and green box) remain conserved at base pair level but can undergo rearrangements (region 2 and 3). The black boxes represent neutrally evolving intervals which may or may not be conserved depending on the divergence time and genomic context. The double headed arrows indicate the alignable subsegements of the homologous sequences.

The rationale behind the local alignment methods is to search for highly similar regions in two sequences, where the regions of similarity are not conserved in order and orientation (regions 2 and 3 in Fig. 1.2). Local alignment methods are useful when long genomic intervals are compared (order and orientation may not necessarily be preserved in long regions).

The global alignment methods are useful to draw an overall map between sequences which are sufficiently small to fulfill the assumption of preserved order and orientation, because they reject those alignments that overlap or crossover (for example, region 2 and 3 in Fig. 1.2 will not be detected by global alignment programs).

Below, the most popular local and global alignment algorithms implemented in PipMaker (BLASTZ) and VISTA (AVID, LAGAN, SLAGAN and MLAGAN) webservers will be introduced and compared to reveal their relative advantages and shortcomings:

BLASTZ (Schwartz et al. 2003) is a local sequence alignment algorithm, and like all sequence alignment programs it follows three major steps. In step one, BLASTZ performs the seeding, i.e. it looks for match of 19 consecutive nucleotides between the sequences to be aligned. Within these 19-mers (1110100110010101111) the 12 positions indicated by 1 should be identical except a single transition event, which is allowed at one of the 12 positions. In step 2, BLASTZ extends the seeds in each direction without allowing gaps, until

the score decreases some given threshold. In step 3, it further extends the alignment, now allowing gaps through dynamic programming. If the distance between two adjacent initially generated BLASTZ sequence alignments is <50 kb in both sequences, the algorithm recursively searches the regions between them by using a more sensitive seeding procedure (e.g. 7-mers exact match) and lower threshold for both gap free alignment and gapped alignment.

The local aligners, like BLASTZ, are highly sensitive when the genomic sequences to be compared are long, unordered and harboring many rearrangements. However, the shortcoming of local aligners is that, they are less specific and can potentially display many false positives (regions which are not true homologs). The working principle of BLASTZ is shown in Fig. 1.3.

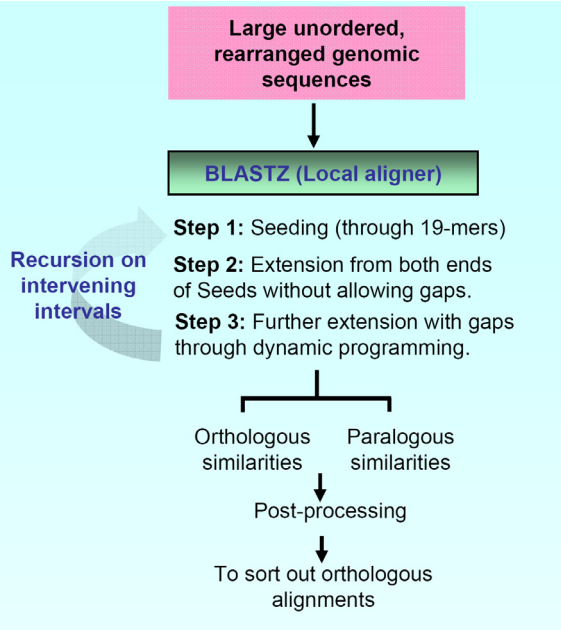


Figure. 1.3. The local alignment algorithms, such as BLASTZ, increase the sensitivity of the alignments at the cost of specificity.

Diagram representing the three step strategy used by BLASTZ to generate a local map between an input sequence data set.

AVID (Bray et al. 2003) is a global aligner and generates a global map by a three step process:

i) It finds the exact matching sequences flanked by mismatches between the input sequences.

ii) AVID generates the anchor set. An anchor set is a collection of non-overlapping non-crossing matches (red matches shown in Fig. 1.4).

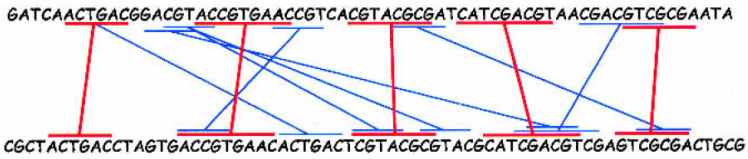


Figure. 1.4. Global alignment algorithms such as AVID are highly specific in detecting the conserved homologous intervals when input sequences are short and underwent no sequence rearrangement events.

Diagram illustrating anchor selection from a set of all possible initially matches that has been localized by the global aligner under the given criteria (adopted from Bray et al. 2003).

The algorithm uses clear matches (outside of known repeats) first, followed by repeat matches to generate a set of anchors. The algorithm then applies step 1 and 2 recursively on the unalignable regions between the adjacent anchors until either no base remains to be aligned or no significant matches can be found.

iii) Sequences that remain to be aligned and that are short enough (< 4 kb) are aligned by using a classical Needleman-Wunsch algorithm. If the sequences are long (>4 kb), the lack of anchors indicates that no significant alignment exists between them.

The AVID algorithm assumes that the two sequences to be aligned, share large blocks of essentially uninterrupted synteny, however, this assumption is incorrect when applied to very large genomic intervals (>8 Mb length in human /mouse comparison). AVID is highly specific in detecting the true homologs, which appear in the same order and orientation between the input sequences, while it is less sensitive than BLASTZ in a sense, that it can miss true positives in a process of rejecting the alignments that overlap or cross-over.

LAGAN (Brudno et al. 2003a) is another global aligner that uses the same three step strategy as discussed above for AVID to generate a global map between the two input sequences. The difference between LAGAN and AVID is that, AVID uses exact matching words to generate the local alignments in step 1, and thus is most suitable for aligning the sequences from closely related species, while LAGAN performs initial local alignments with

techniques designed to work well on distantly as well as closely related organisms. It uses the highly sensitive CHAOS program, which detects the local alignments by using short inexact words, instead of using long exact matches. Another advantage of LAGAN over AVID is that, LAGAN can align larger sequences (> 2 Mb) owing to considerable lower memory requirements. Like AVID, LAGAN cannot detect genomically rearranged conserved fragments. The working principal of a typical global aligner (AVID/LAGAN) is shown in Fig. 1.5.

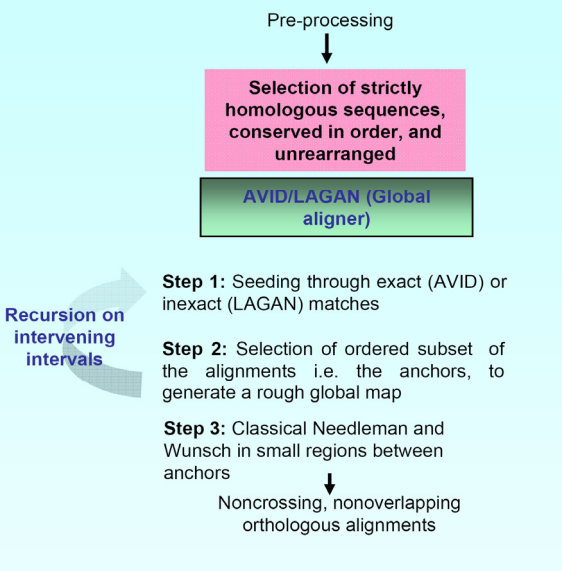


Figure. 1.5. Seeding through CHOAS program makes LAGAN to perform slightly better than AVID.

Diagram representing the three step strategy used by AVID/LAGAN to generate a global map between an input sequence data set.

MLAGAN (Brudno et al. 2003a) is an extension of LAGAN and it enables the user to perform multiple alignments of large genomic sequences. MLAGAN is based on a progressive alignment strategy. A multiple alignment of k sequences is generated in k-1 pairwise alignment steps, where in each step two closest sequences or intermediate multiple alignments are aligned using the phylogenetic tree as a guide.

Shuffle-LAGAN (Brudno et al. 2003b) is a glocal alignment algorithm, which is a hybrid of global and local methods, based on the CHAOS local alignment algorithm and the LAGAN global aligner. SLAGAN is able to align long genomic sequences while detecting DNA rearrangement events (translocations, inversions and duplications) and simple edits

(changes at single base pair level). SLAGAN has better sensitivity and similar specificity than LAGAN, while in comparison to local aligner BLASTZ, it has better specificity but less sensitivity.

1.5 Motif finding

The gene regulatory function of an enhancer element is conferred by its interaction with *trans*-acting regulatory factors. The transcription factor binding sites (TFBSs) are short sequence motifs, usually consisting of 6-12 base-pairs, often arranged as "high-order site clusters" within a typical eukaryotic enhancer. Such clusters can be homotypic, containing multiple sites for a particular transcriptional factor or heterotypic, containing one or more binding sites for multiple transcription factors (Bulyk 2003). Traditionally, these TFBSs were determined by labor-intensive wet-lab techniques such as DNAase footprinting and gel-shift assays and several online databases like TRANSFAC (http://www.biobase.de) have been constructed to store the binding site information for ~500 vertebrate-specific TFs (transcription factors). This wealth of TFBSs information for vertebrate-specific TFs provides unprecedented opportunity to improve our understanding about *cis*-acting regulatory components of the vertebrate genome broadly in two ways:

i) To predict and prioritize candidate regulatory regions from bulk of "dark matter" of vertebrate genome for functional analysis depending solely on their in-silico TF binding properties.

ii) To determine the gene regulatory potential of an enhancer element experimentally and to focus subsequently, the computational prediction and prioritization of short functionally relevant motifs (TFBSs) for further experimental analysis.

A

Pos	1	2	3	4	5	6	7	8	9	10
A	0	0	1	25	19	7	1	2	2	0
C	0	0	0	0	13	1	2	17	35	36
G	38	38	37	13	1	3	2	0	0	0
T	0	0	0	0	5	27	33	19	1	2

B

Pos	1	2	3	4	5	6	7	8	9	10
A	-2.8	-2.8	-2.1	1.3	0.9	-0.4	-2.1	-1.6	-1.6	-2.8
C	-2.8	-2.8	-2.8	-2.8	0.4	-2.1	-1.6	0.8	1.7	1.8
G	1.8	1.8	1.8	0.4	-2.1	-1.3	-1.6	-2.8	-2.8	-2.8
T	-2.8	-2.8	-2.8	-2.8	-0.8	-1.4	1.7	0.9	-2.1	-1.6

Figure. 1.6. Currently position weight matrices (PWM) are being used to represent the binding profiles of experimentally verified transcription factors in databases like TRANSFAC and JASPAR.

Representation of NF-kB binding sites. (A) Count matrix for NF-kB resulting from an alignment of 38 experimentally verified functional binding sites. (B) Position weight matrix (PWM) for NF-kB, which is the logarithm of the frequency counts divided by expected counts. Adopted from (Sandelin et al. 2004).

To model and predict individual TFBSs within a target sequence is proved to be much harder than predicting genes, the intrinsic difficulty is being that TFBSs are in general very short and often degenerate in sequence. Several strategies are used to classify TFBSs in different databases in a way to make them maximally informative for various TFBSs prediction algorithms. The most inflexible way is to use a single unambiguous sequence to categorize a specific binding site (for example TATAA). Alternatively, the set of experimentally identified TFBSs for a given TF can be aligned to generate consensus binding sequence which can incorporate the ambiguous positions. Although the degenerate consensus sites are still frequently used to predict the binding specificities of TFs but they do not contain the precise information about the relative likelihood of observing the alternative nucleotides at the various positions of a given TFBS. Recently, the consensus sequence approach has been superseded by the matrix approach, in which a set of known binding sites can be aligned and the frequency of individual nucleotides at each position is counted to generate a count matrix (Fig. 1.6A). The count matrix is converted to a logarithmic scale for computational analysis. The resulting log-converted matrices are known as position weight matrices (PWM), and can account the observed frequency of tolerated sequence variations at each nucleotide position within a consensus TFBS, gives a quantitative score and reflects the actual binding specificity of the given TF (Fig. 1.6B). Currently, PWMs are used to represent the binding profiles of experimentally verified TFs in databases like TRANSFAC and JASPAR. The use of PWM's binding profiles for TFBSs predictions on a single sequence is problematic, as TFs bind to short degenerate motifs and such motifs can occur very frequently in the genome just by chance, which can result in large numbers of false positive predictions with no biological significance. For instance, the unambiguous sequence TATA is expected once every 1024 bp by chance, which predicts 30 million potential binding sites in the mammalian genome. Several strategies have therefore been developed to reduce the false-positive rate. These include combining predictions with gene expression data or use of prior knowledge of gene-coregulation. Another approach is to take advantage of the fact that genes are often regulated by multiple TFs, therefore potential TFBSs tend to be clustered or adjacent to each other. Some researchers have tried to couple the genome alignment tools with PWMs to create a system that is sensitive and easy to use (Loots and Ovcharenko 2004; Sandelin et al. 2004). The technique has been known as phylogenetic footpriniting, a term inspired by the wet-lab technique DNAse footprinting. The rationale behind this approach is that binding site footprints (6 bp or more) found in human sequence that are also found at corresponding orthologous positions in mouse or other mammalian and non-mammalian sequences are far more likely to be real than those found only in humans. On average, the phylogenetic

footprinting improved the selectivity of TFBS prediction by 85% compared to using a matrix model alone (Lenhard et al. 2003) dramatically reducing the number of false positive matches.

Below, the two most popular freely available phylogenetic footprinting web tools based on TFBS prediction will be discussed.

The **ConSite** algorithm (Sandelin et al. 2004) searches for the conserved TFBS motifs between the input orthologous intervals by proceeding through the following steps:

i) It aligns the submitted pair of orthologous sequences by using the ORCA aligner which is a global alignment program.

ii) It collapses the gaps in the alignment and calculates a separate conservation profile for each orthologous sequence.

iii) It scans each of the sequences for PWMs from the JASPAR database.

iv) It performs filtering on the initial sets of TFBSs using phylogenetic footprinting and presents those conserved TFBSs which score higher than the user defined relative matrix score threshold.

rVISTA analysis (Loots and Ovcharenko 2004) proceeds in four main steps:

i) It performs the alignment between the input data set by using the local alignment program BLASTZ.

ii) It detects TFBS matches in each individual sequence using PWMs from the TRANSFAC database.

iii) It identifies pairs of locally aligned TFBSs.

iv) It selects TFBSs present in regions of high DNA conservation (80% identity, 20bp sliding window) for presentation.

1.6 Innovation of limbs provides an insight into, how the vertebrates achieved morphological complexity during their evolutionary history

Vertebrates are part of the phylum Chordata. Two major phyla, Ambulacraria and Chordata, constitute the supertaxon of deuterostomes. Recent phylogenetic reconstructions based on biochemical data (DNA/protein) placed the hemichordates as a sister group of echinoderms. Together these two form a discrete group, the Ambulacraria. The tunicates, cephalochordates and vertebrates together constitute, the Chordata (Fig. 1.7).

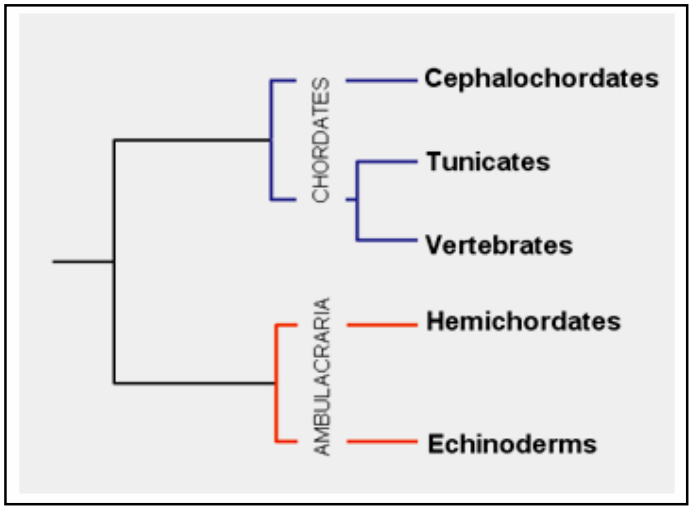


Figure. 1.7 Proposed phylogeny of the deuterostomes

Hemichordates (acorn worms) are worm-like solitary animals, a few centimeter to two meter in length, with up-to several hundred pairs of gill slits. The body constitutes three parts (prosome, mesosome and metasome) each with coelomic cavity or paired cavities. The dorsoventral arrangement of organs in hemichordates is like in protostomes: ventral nervous system, ventral musculature and dorsal heart.

Echinoderms include a diverse group of marine animals. Despite their unique penta-radial symmetry as adults, these animals begin life with bilateral symmetry as larvae. Another important feature of echinoderms is that most have tube feet for locomotion and holding onto prey.

The gill slits might be an ancestral deuterostome trait (Gerhart et al. 2005) as they persist in both Ambulacraria (hemichordates) and Chordata (tunicates, cephalochordates, and vertebrates), while considerable changes had occurred in the chordate line, particularly the centralization of the nervous system, the organization of the notochord, and the inversion of the dorsoventral axis (Fig. 1.8).

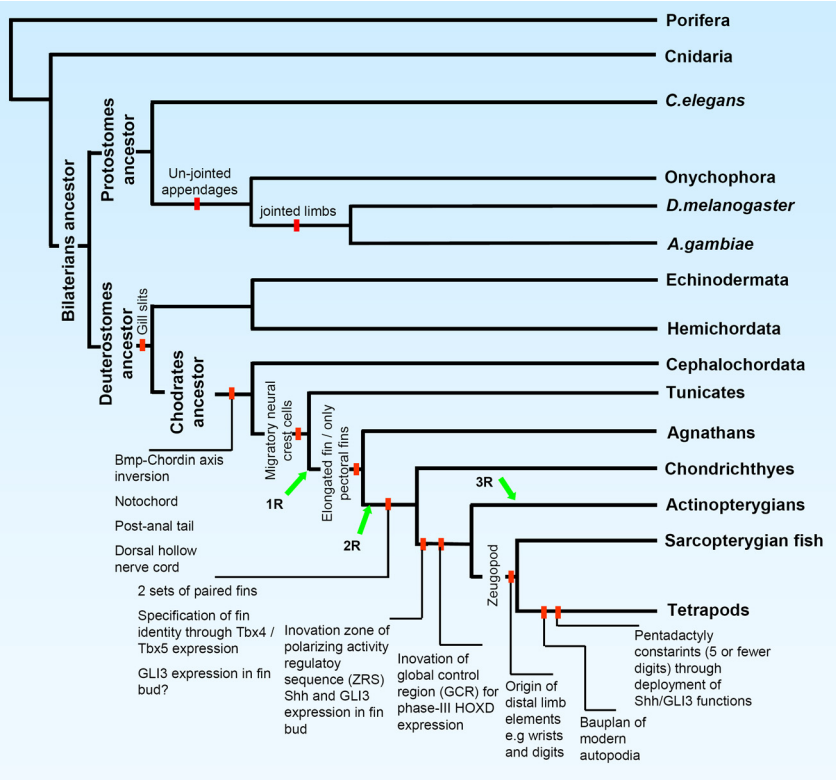


Figure. 1.8. Evolutionary history of bilaterians in context of approximate time points of morphological innovations leading to greater complexity in the body plans of modern vertebrates.

Diagram illustrating the approximate history of major molecular and morphological innovations along with the origin of limb elements during the evolutionary history of protostome and deuterostome lineages. IR, 2R and 3R refers to proposed genome duplication events experienced by the vertebrate lineage. Note: description and references of the evolutionary events shown in this Fig. are given below in the text.

In addition to the synapomorphies discussed above, uniting the various groups of chordate lineage another important innovation that took place in the ancestor of tunicates and vertebrates after the divergence of amphioxus (cephalochordata) (Fig. 1.8) is the capacity of neural tube to generate migratory pigment cells (in tunicates) (Jeffery et al. 2004) which probably near the base of vertebrate radiation attained additional properties including pluripotency, delamination, migration, and carriage of anteroposterior positional information. These novel properties were essential for the evolution of patterned craniofacial structure, which is a defining vertebrate synapomorphy.

Another important feature that differentiates vertebrates from other chordate lineages is that they are actively feeding and predatory while their closest invertebrate relatives, tunicates and cephalochordates, are filter feeding and sessile. However, the ascidians (tunicates) have a free swimming larval stage in which the tadpole swims with the help of a propulsive tail (Fig.

1.9A). This motile larva subsequently glues its face to a rock and undergoes metamorphoses by resorbing its tail, rearranging its nervous system, and developing a pair of siphons to filter out the algae from the sea for the rest of its life (Fig. 1.9B).

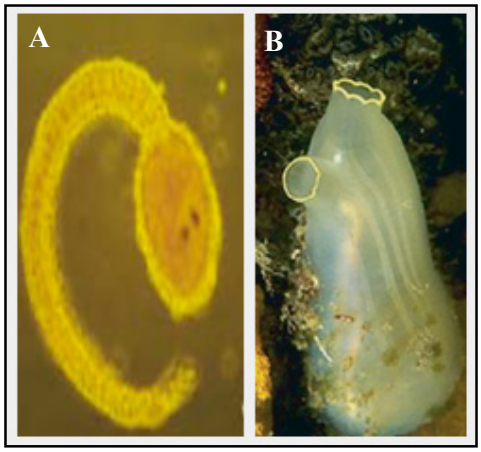


Figure. 1.9. Variations in the life cycle of tunicates, a sister group of vertebrates, provides an insight into how the free swimming mode of locomotion might have first appeared in vertebrates.

(A) free living motile larva and (B) sessile adult stage of *C.intestinalis*. Adopted and modified from (Canestro et al. 2003).

The variation in the life cycle of ascidians led classical anatomists to suggest that freely swimming mode of locomotion in vertebrates might be evolved by retaining the motile larval form of ancestors (tunicates) as the adult form of descendants (vertebrates), a phenomenon known as paedomorphosis.

The primitive fish-like vertebrates (Fig. 1.10) in Ediacaran and early Cambrian seas might have used purely undulatory motions, by side to side vibration of the body caused by the action of myotomes to move.

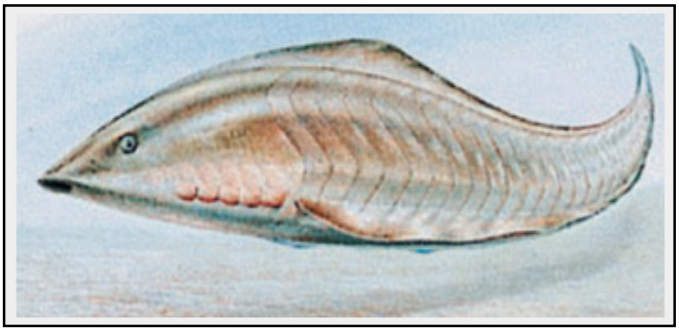


Figure. 1.10. In primitive vertebrates the swimming activity was independent of appendicular architecture.

Reconstruction of the early Cambrian craniate Myllokunmingia using side to side body vibrations for motion. Adopted and modified from (Shimeld and Holland 2000).

The next remarkable innovation in the history of vertebrate evolution was the origin of body appendages to facilitate the feeding and locomotion in Ordovician seas (Shubin et al. 1997). This innovation first took place in jawless vertebrates, with some having had ribbon-shaped fins extending laterally along the body wall, others having had paired pectoral fins only, protruding immediately posterior to the head region (Fig. 1.11A & B).

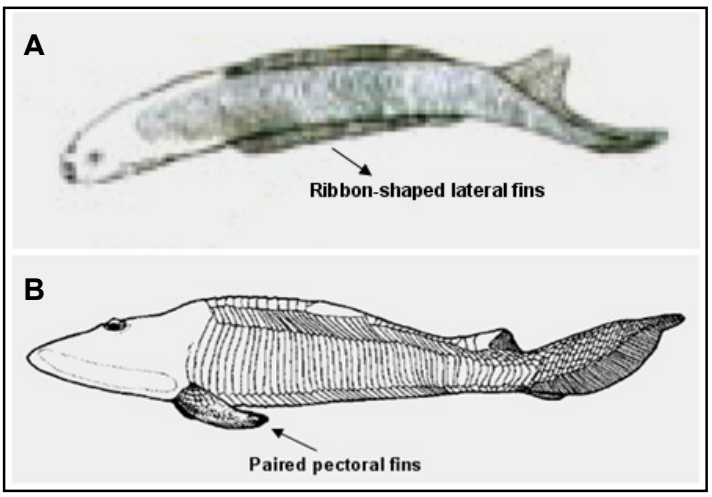


Figure. 1.11. Body appendages serving locomotion and feeding first appeared in jawless vertebrates.

(A) Elongate fins and (B) paired pectoral fins in Silurian jawless fish (adopted and modified from Shubin et al. 1997; Coates and Cohn 1998).

The two sets of paired appendages are the synapomorphy of ganthostomes (Coates and Cohn 1998; Shubin et al. 1997) and first made their appearance in the Silurian era in the form of pectoral and pelvic fins. Since their origin, the pectoral and pelvic fins are believed to be evolved in parallel (serially homologous) in almost all major ganthostome clades (Shubin et al. 1997). In all jawed vertebrates the pectoral fin skeletons share the following ancient structures: A series of radials (internal skeleton support) articulate with the girdle, the posterior most radial further articulates with secondary radials along its anterior edge. This complex posterior radial is known as metapterygium (Fig. 1.12) which was lost in one major vertebrate group, the teleosts (for example, zebrafish, and Fugu). Soft fin rays, constructed of collagen-like protein, constitute the periphery of both pectoral and pelvic fins. In bony fishes these soft fin rays incorporate dermal bony rays (lepidotrichia) (Fig. 1.12). Pelvic fins in contrast to pectoral fins are small and anatomically more simple, and some investigators doubt the serial homology of pelvic and pectoral appendages (Coates and Cohn 1998).

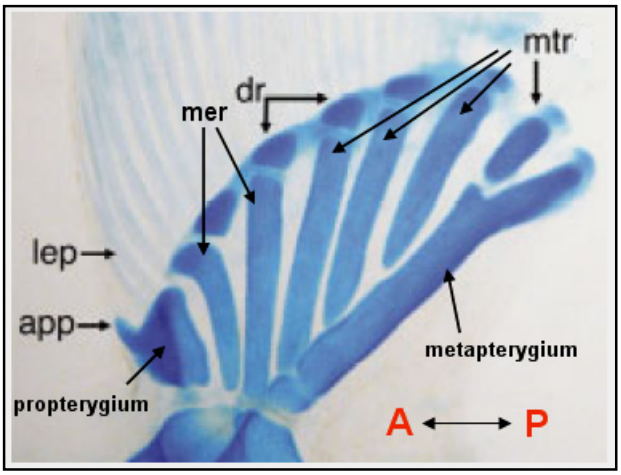


Figure. 1.12. The skeletal elements of paddlefish pectoral fin show similarities to tetrapod limb skeleton.

Pectoral fin of *Polyodon spathula* (paddlefish) containing elements considered homologous to both the fin radials of teleosts and the limb skeleton of tetrapods. mtr; metapterygial radials, dr; distal radials, mer; mesopterygial radials, lep; lepidotrichia, app; anterior process of the propterygium. Adopted and modified from (Davis et al. 2004).

For the fin to limb transition and the origin of digits in tetrapods, Shubin and Alberch proposed a model to define homologous regions of tetrapod limbs and paired fins of Sarcopterygians (lobe fin fish, the sister group of tetrapods) (Shubin et al. 1997). According to this model the proximal parts (humerus/femur, radius/tibia, ulna/fibula) are homologous in tetrapod limbs and Sarcopterygians (Panderichthyes) fins (Fig. 1.13A & B). In order to explain the evolutionary origin of distal parts in tetrapod limb, Shubin and Alberch proposed that during limb development the distal row of carpal/tarsals arise from digit arch (Fig. 1.13B). The evolution of digital arch itself resulted from a bending of the metaptergial axis (main stem of branched pattern formed by fin metapterygia and associated radials) of the sarcopterygian fin. This axis is believed to have developed from proximal to distal. According to this model, what was originally distal in sarcopterygian fin is now anterior in tetrapods. Thus, during fin to limb transition there has been a turning of the proximodistal axis towards the anterior. The anterior digits, therefore, correspond to the most distal region of the limb.

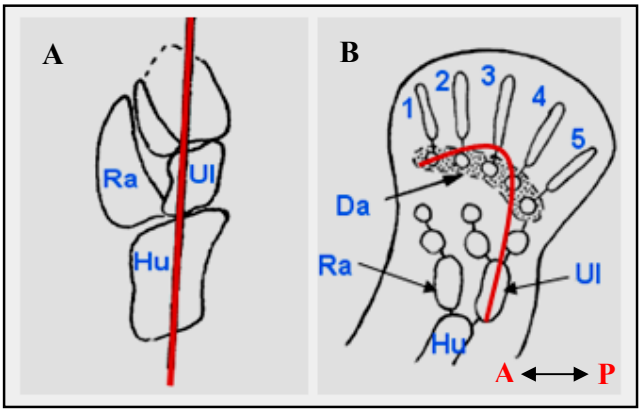


Figure. 1.13. The digit arch model describes how the distal limb elements in tetrapods might have evolved through bending of the ancient metapterygial axis.

(A) Panderichthyid (sister group of tetrapods) fin with hypothetical metapterygial axis (marked red) running straight from proximal to distal end of fin and radials branching preaxially. (B) in tetrapods there has been a turning of metapterygial axis (marked red) towards the anterior of the proximodistal axis. The digit arch and the digits lie on the postaxial side. Da; digit arch, Ra; radius, Ul; ulna, Hu; humerus. Adopted and modified from (Hinchliffe 2002).

The concept of the metapterygial axis further suggests a posterior dominance during vertebrate limb development (Coates and Cohn 1998) and thus the branching of the cartilaginous elements of the radius and ulna from the more proximal humerus is proposed to be asymmetrical. The anterior side (radial/tibial) does not branch in general, while the posterior (ulnar/fibular) may branch. The majority of the distal elements are hypothesized to arise from the posterior region of the limb. The digital arch model proposes that metacarpals/metatarsals arise by bifurcation from carpals/tarsals, whereas the phalanges arise by segmentation from more proximal elements.

1.6.1 Fin to limb transition involved cis-regulatory networks

Fin to limb transition was one of the most important macroevolutionary changes that took place in the history of vertebrate evolution. Molecular studies of the main fin axis in cartilaginous fish, teleost fin and tetrapod limb development have uncovered striking conservation of the genetic control of pattern formation along the fin/limb axis (Sordino et al. 1995; Tanaka et al. 2002; Vanderlaan et al. 2005). The major morphological innovations that took place during fin to limb transition are

i) the identity of each pair of appendages,

ii) freeing the fins from the body axis and establishing a separate limb axis,

iii) innovation of the autopod with digits in tetrapods.

In tetrapods and teleosts, the identity of each pair of appendages is determined by the expression of *Tbx5* and *Tbx4* in the anterior and posterior appendages, respectively. Amphioxus (cephalochordata), the closest relative of vertebrates, has only one T-box gene, AmphiTbx4/5 (Ruvinsky et al. 2000), which by duplication (deep in the vertebrate linage) gave birth to vertebrate *Tbx4* and *Tbx5*. This innovation of a genetic circuit for limb specification by recruiting the expression of the T-box gene family members to vertebrate paired appendages took place as early as before the divergence of cartilaginous fish lineage (Fig. 1.8). The evidence for this early acquisition of limb specification is provided by the observation that the expression domains of *Tbx5* and *Tbx4* genes in dogfish embryos are consistent with their expression patterns within the anterior and posterior appendages of teleosts and tetrapods (Tanaka et al. 2002).

In many cartilaginous fish, the metapterygium (the main long bone of the fin) develops parallel to the body axis: Freeing the fin from the body and establishing a separate proximo-distal limb axis has been correlated with the acquisition of *Shh* expression in the appendages (Tanaka et al. 2002). The recruitment of *Shh* expression to the fin bud took place after the divergence of the cartilaginous fish lineage (Fig. 1.8), as Tanaka et al. (2002) could not detect SHH in dogfish fin buds. In contrast, in teleosts (zebrafish) and mammals (mouse), *Shh* is expressed in the posterior region of fin/limb buds and these groups have paired appendages with a separate proximo-distal axis. Recently, the discovery of a limb/fin specific *Shh* enhancer (*zone of polarizing activity regulatory sequence,* ZRS) (Fig. 1.14) suggested that fin/limb specific polarized expression of *Shh* in vertebrates might have been achieved through the innovation of this regulatory element after the divergence of cartilaginous fish lineage and before the splitting of tetrapod-teleost lineage (Lettice et al. 2003) (Fig. 1.8). The sequence of ZRS and even its relative position with respect to the *Shh* transcription initiation site is highly conserved in tetrapods and teleosts and it has also been shown that activity of this element is exclusive for limb specific functions of *Shh* as its deletion results in complete loss of *Shh* expression only in the limb bud associated with degeneration of distal limb skeletal elements (Sagai et al. 2005).

The molecular investigations with dogfish embryos (Tanaka et al. 2002) further suggest acquisition of the expression of *En1*, *Bmp4* and *dHand* as early limb patterning regulators to the fin bud, at least before the divergence of the cartilaginous fish lineage.

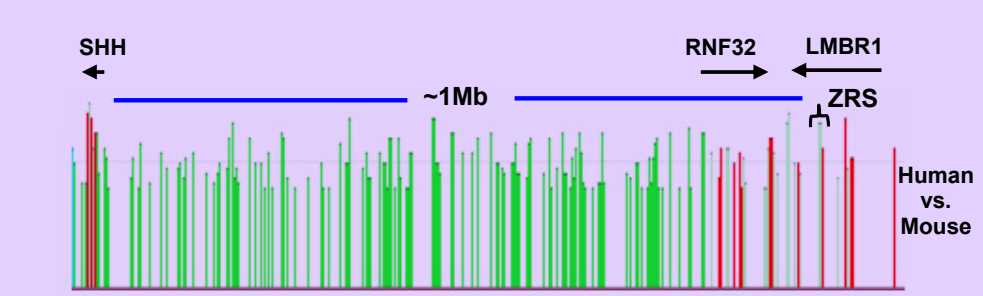


Figure. 1.14. A distant, evolutionarily conserved enhancer region (ZRS) controls *SHH* expression along the posterior edge of the emerging limbs from a distance of ~1Mb.

Diagram illustrating a human vs. mouse sequence identity plot of the chromosomal segment containing the *SHH* locus and its regulatory sequences. The green and red bars represent the conserved non-coding and coding intervals respectively. Above the plot, the arrows depict the genes within the region and their direction of transcription. The ZRS (*zone of polarizing activity regulatory sequence*) is shown with a bracket symbol within intron-5 of the *LMBR1* gene. The blue bar above the graph depicts the approximate distance between ZRS and the *SHH* transcription initiation site.

The fins of teleost fish have been under intense molecular investigations, to draw a map of molecular events which ultimately led to fin to limb transition. Among the genes shown to be expressed in both teleost fins and tetrapod limbs are *Shh, Ptc1, Bmp4, Fgf8, Dlx*, *Fgfr, Bmp2, Hoxa, Hoxd, Hoxc6, Msx, En1, Sal1, Gli3,* and *dHand*. The similarities in molecular architecture are not confined just to simple presence and absence of gene expression, instead the precise spatial relationships in early fin development bear striking resemblance to those found in tetrapod limb development.

If fish fin and tetrapod limb early development involves this strikingly similar setup of signaling networks, then how can the extreme morphological differences be achieved. The most obvious difference between the tetrapod limb and teleost fin is found distally. From the available comparative studies it appears that molecular patterning of proximal elements (stylopod and zeugopod) is achieved by the same mechanism in fish and tetrapods, but important differences in gene expression patterns occur later, during the development of distal elements. For example one major molecular difference in patterning of distal limb and fin component is *Shh* expression in the posterior region of the limb bud mesenchyme, controlling anterorposterior patterning of the limb all the way to digit development, while in zebrafish *Shh* expression in the fin bud disappears prior to ray formation. This early loss of *Shh* expression in the posterior of fin bud is associated with the lack of a third phase of *Hox* gene expression in the fin bud. Recent data suggests that differential expression of the 5´genes of the *HoxD* cluster during the last phase of limb development may have been crucial to the

invention of distal structures (autopod with digits) in tetrapods (Sordino et al. 1995). In today's limbed vertebrates, during a first phase of Hox expression, genes from the middle of the *HoxD* cluster, i.e. *Hoxd9* and *Hoxd10,* are expressed in proximal embryonic limb tissues, which will become the stylopod (Fig. 1.15). During the phase II pattern of *Hox* gene expression, the more 5` *Hoxd* genes, *Hoxd11* to *Hoxd13* are also expressed, with Hoxd9 through Hoxd13 expression domains occupying the posterior of the limb bud, while Hoxd9 is expressed both anteriorly and posteriorly (Fig. 1.15). Acquisition of the phase II expression pattern of *Hox* genes led to zeugopod formation. Zebrafish and tetrapods share the phase I and II pattern of Hox expression in their appendages. Thus, both groups form stylopod and zeugopod. However a 3rd phase of Hox gene expression is unique to tetrapods and is not found in zebrafish. During phase III, the spatial order of posterior *Hoxd* genes is reversed compared to their normal expression in trunk and zeugopod (Sordino et al. 1995) and also their quantitative collinear manner of expression is inverted during this phase (Zakany et al. 2004). Normally 5´end *Hox* genes are expressed at higher levels and their 3´located neighbors are expressed at progressively lower level. In the autopod, *Hoxd13* is expressed most anteriorly and at the highest level while *Hoxd12*, *Hoxd11* and *Hoxd10* are expressed at progressively lower levels and stepwise in more posteriorly restricted domains defining a "HOX-code" (Fig. 1.15). This reversal in spatial and quantitative polarity of posterior *Hoxd* gene expression is considered to be correlated with the bending of the metapterygium axis towards the anterior of the proximo-distal axis during limb evolution (Shubin et al. 1997) which led to the origin of the autopod where the anterior digits correspond to the most distal region of the limb (Fig. 1.13B). Phase III pattern of Hox expression is in fact altogether absent from zebrafish appendage, suggesting that this pattern might be an apomorphy for tetrapods and that these regulatory changes in *Hox* expression may underlie the origin of the autopod.

However, in contrast to comparative molecular and genetic data from representative members of teleost (zebrafish) and terrestrial vertebrate lineages (mouse), the data from fossils support the notion that the unique features of tetrapod limbs (autopodia) were assembled over evolutionary time in the paired fins of fish and thus are not an evolutionary novelty of tetrapods (Shubin et al. 2006). If this were the case, than phase III of *Hox* gene expression might have been acquired by the bony fishes before terrestrial invasion. In agreement with the fossil data, recently, the expression analysis of limb/fin developmental regulators in paddlefish *Polydon spathula* (a basal actinopterygian) revealed a late phase, inverted collinear expression of 5 *HoxD* genes (Davis et al. 2007). This conservation of late

phase *HoxD* genes expression suggests that the pattern arose before the divergence of tetrapods and actinopterygian fish (Fig. 1.8) and was subsequently lost in teleosts.

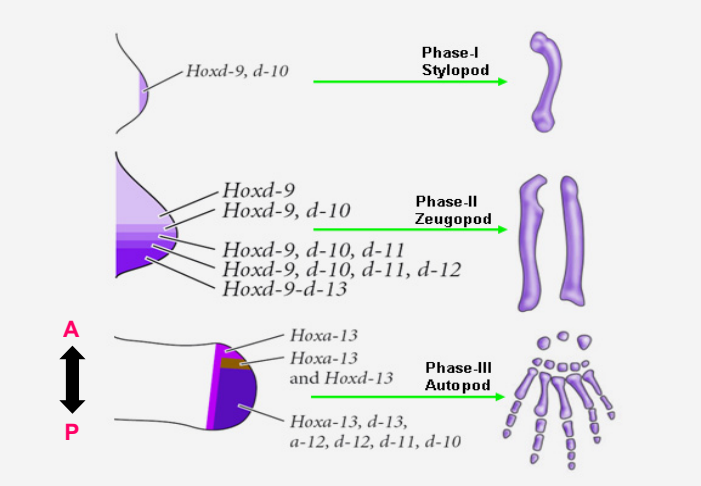


Figure. 1.15. Changes in *HOX* gene expression during the formation of tetrapod limb

During phase I, *Hoxd9* and *Hoxd10* are expressed in the newly formed limb bud which leads to the formation of the proximal element, i.e. stylopod. During phase II there is nested expression of *HoxD* genes. *Hoxd9* through *Hoxd13* are expressed in the posterior of the limb bud, while *Hoxd9* is expressed both anteriorly and posteriorly. This nested expression pattern leads to the formation of intermediate elements, i.e. the zeugopod. During phase III, the inversion of *HoxD* expression leads to the formation of autopod elements. Adopted and modified from (Shubin et al. 1997).

The evolution of novel distal elements of tetrapod limbs (digits) did not involve the origin of novel developmental regulators to govern the anterior turning of the ancient proximodistal axis, instead a part of the already existing ancient genetic circuit was recruited to perform this novel task, through inversion of its ancient spatial, temporal, and quantitative collinearity.

The collinearity in the expression pattern of *HoxD* genes along the main embryonic axis and proximal elements of tetrapod appendages is generally regulated by enhancers positioned within the *HoxD* cluster (Spitz et al. 2003), which govern the sequential deployment of *HoxD* genes in both space and time, with 3´ (teleomeric) situated genes (for example, *HoxD9*) expressed first and anteriorly, while centromeric neighbors (for example, *HoxD13)* expressed later and in posterior tissues.

The inversion in the expression behavior of posterior *HoxD* genes in the distal limb results from the of innovation of a digit enhancer (Kmita et al. 2002; Peichel et al. 1996; Spitz et al. 2003). This element, a Global Control Region (GCR) of ~40 kb, which exert its effect on 5´

HoxD genes (*HoxD10-13*) from the distance of ~250 kb (telomeric), has been localized by Spitz et al (2003) with a combination of a transgenic mice assays and computational tools (Fig. 1.16). The parallel occurrence of phase III pattern of *HOX* expression in mouse and paddlefish suggests that the digit enhancer has been innovated in the common ancestors of tetrapods and actinopterygian fish (Fig. 1.18) and was subsequently lost in teleosts as multiple digit enhancers embedded within the GCR, which are necessary and sufficient to recapitulate expression in digit development appear highly conserved in human-mouse but not in human-pufferfish sequence comparison (Fig. 1.16). Whether or not the interval spanned by the GCR in mammals is conserved in paddlefish, as is the late-phase *Hox* expression must await the genomic sequence of paddlefish for comparison.

Thus both the fossil record and molecular data suggest that genetic and developmental basis of autopod evolution are primitive to tetrapods and the origin of digits entailed the redeployment of ancient patterns of gene activity (Davis et al. 2007).

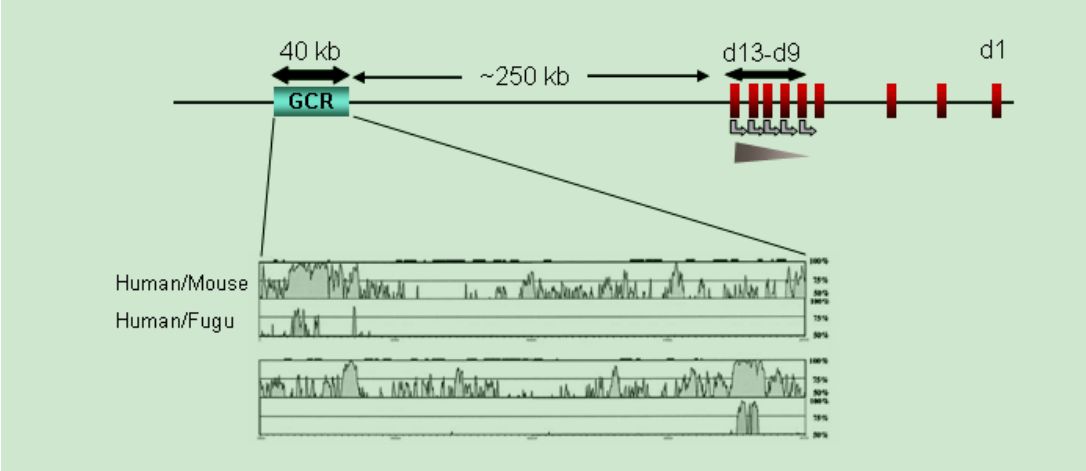


Figure. 1.16. Putative enhancer elements embedded within the GCR influence the posterior *HOXD* genes expression specifically during digit development.

Architecture of the human *HOXD* locus, with posterior *HOXD* genes (d13-d9) under spatial, temporal and quantitative (triangle underneath the d13-10 genes) influence of GCR (global control region) during phase III of *HOX* gene expression. Vista plot depicting the sequence conservation at GCR locus between human-mouse and human-Fugu intervals. Adopted and modified from (Spitz et al. 2003).

1.7 *Cis*-regulatory modules in human disease

The cis-acting regulatory network of an early developmental regulator typically contains several enhancers (activators/silencers), which can be located 5` and 3` and also within intronic intervals. Each enhancer usually controls expression in a subset of tissues in which the associated gene is normally expressed and the complex expression patterns of

developmentally important genes is usually mediated by many enhancers accumulatively (Nobrega et al. 2003). A typical enhancer element can be of ~500 bp in length and potentially harbors the binding sites for many transcription factors. Combinatorial effects of these transcription factors dictate the precise activity and target of the associated enhancer element (Remenyi et al. 2004). The enhancers can impose their effects on the associated target genes over a distance of roughly 10 kb in *Drosophila* and 100 kb-1000 kb in humans. Sequence variations within the cis-acting regulatory elements can have drastic consequences on the expression pattern of the associated genes and can thus cause morphological, physiological, or behavioral modifications (Anand et al. 2003; Shashikant et al. 1998). Consequently, alterations within *cis*-acting regulatory elements might also lead to human developmental disorders. Traditionally, the search for mutations causing human developmental disorders has focused on those that change the coding sequence of a protein. From 10-15% of patients with abnormal Mendelian phenotypes depict no alterations within the coding sequence of their candidate gene. In some of these cases it should be expected that perhaps the unknown causative mutation resides within the non-coding regions, presumably in regulatory intervals of the genome. Only 1% of mutations which affect the gene functions through alteration of the proper gene expression have been identified and localized predominantly within minimal promoter regions (Stenson et al. 2003). Enhancers scattering over a large distance from the associated genes could potentially harbor disease associated mutations (Emison et al. 2005; Lettice et al. 2003). However, the identification of disease associated mutations within remote (intronic/intergenic) enhancers has largely been overlooked in the past because of the fact that we currently posses few insight into how to identify functional sequence outside of the coding exons. Consequently this results in a gross underestimate of disease causing non-coding mutations.

Transcriptional regulation networks can be disrupted either by physical dissociation of a gene from part or all of its regulatory network or through the inactivation of one or more cis-regulatory modules by mutation or deletion. In the former case, the chromosomal rearrangements are causative. There are many well documented examples of disease causing translocation breakpoints mapping outside the affected gene (Table 1). In these cases, known as “position effect”, the chromosomal rearrangement has displaced the gene relative to associated cis-regulatory modules, which leads to inappropriate gene expression and thus a pathogenic effect (Table 1). For example, position effects have been detected underlying cases of aniridia, a congenital malformation of the eye caused by haploinsufficiency of the *PAX6* gene. Aniridia patients with translocation breakpoints mapping 100-125 kb downstream of the *PAX6* gene have been described (Kleinjan and van Heyningen 1998). Mutations in the

GLI3 gene lead to the human Greig Cephalopolysyndactyly syndrome (GCPS) (Wild et al. 1997). In one of the GCPS patients the phenotype has been associated with a position effect caused by a translocation break 10 kb downstream of the last exon of *GLI3*. In the mouse, the add (anterior digit deformity) phenotype is caused by a transgene integration at ~40 kb upstream of *Gli3,* representing a putative murine *Gli3* position effect (van der Hoeven et al. 1993). Further examples of position effects in human diseases are listed in Table 1.

Point mutations within enhancer elements can disrupt its *cis*-acting regulatory influence on the associated gene as documented for preaxial polydactyly (PPD). In PPD patients, mutations have been mapped within a *Shh*-associated remote enhancer element (ZRS) residing ~1 Mb 5′ of *Shh* within the intron-5 of the *LMBR1* gene (Lettice et al. 2003) (Fig. 1.14). Similarly a single nucleotide change in an enhancer element within intron-1 of the *RET* gene has been associated with Hirschsprung disease (HSCR) (Emison et al. 2005).

Table 1. Genes affected by position effects in human diseases

Gene	Disease	Distance of furthest breakpoint (kb)
FOXC1	Glaucoma/autosomal dominant iridogoniodysgenesis	25/1200
FOXC2	Lymphedema distichiasis	120
FOXL2	Blepharophimosis/ptosis/epicanthus inverses syndrome	170
FSHD	Facioscapulohumeral dystrophy	100
GLI3	Greig cephalopolysyndactyly syndrome	10
PAX6	Aniridia	125
PITX2	Rieger syndrome	90
POU3F4	X-linked deafness	900
SALL1	Townes-Brocks syndrome	180
SHH	Preaxial polydactyly	1000
SHH	Holoprosencephaly	265
SIX3	Holoprosencephaly	200
SOX9	Campomelic displasia	850
SRY	Sex reversal	3
TWIST	Saethre-Chotzen syndrome	260

Adopted and modified from (Kleinjan and van Heyningen 2005).

1.8 The *GLI* gene family, key developmental regulators

The important components of the hedgehog (Hh) signaling pathway were first elucidated in *Drosophila.* It is now evident that these components remained conserved during millions of

years of separate evolution since *Drosophila* and humans shared the last common ancestor (Fig.1.17)

In *Drosophila*, the Hh signaling pathway is required in patterning of many processes during development. Hh signals are received through interaction of two transmembrane proteins, patched (Ptc) and smoothened (Smo).

In the absence of Hh protein, the transmembrane protein Smo is inhibited by Ptc, thus permitting proteolytic cleavage of Ci into a small N-terminal 75kDa fragment, which then represses the target genes.

Conversely, the binding of Hh to Ptc relieves Smo, which results in the inhibition of Ci-processing and subsequent translocation of full length 155 kDa protein to the nucleus activating transcription from target genes (Fig. 1.17A).

Thus Ci is a dual nature transcriptional regulator which can either repress or activate the transcription depending on Smo activation by Hh.

Following genome amplification events early in evolution (Fig. 1.8) vertebrates have at least three *Ci* homologs, *Gli1, Gli2* and *Gli3*, that encode transcription factors with five tandem C2-H2 type zinc fingers linked by histidine-cysteine bridge sequence (Ruppert et al. 1988).

The three vertebrate Gli proteins have subdivided the different features of Ci functions: Gli3 has retained the highest functional similarity to Ci. Like Ci, it undergoes proteolytic processing to generate a truncated N-terminal protein, which accumulates in the nucleus to repress the target genes (Fig. 1.17B). Primarily, Gli3 functions as repressor of transcription, but subject to Shh signaling, it may also act as activator of Hh target genes. Gli2 like Gli3 can be proteolytically processed and seems to have transcriptional activator and repressor functions while Gli1 does not appear to undergo proteolytic cleavage, and there is no evidence for its repressor activity.

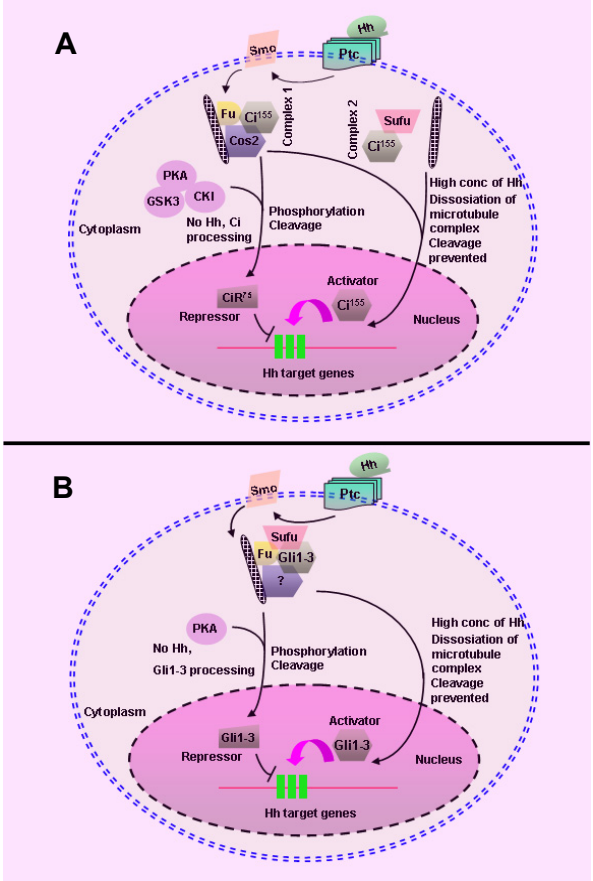


Figure. 1.17. The hedgehog signaling pathway is highly conserved among insects and vertebrates

Schematic representation of hedgehog signalling pathways. **A**) *Drosophila*, B) vertebrates. Ptc, Patched; Smo, smoothened; Cos2, Costal-2; Fu, fused; Sufu, Suppressor of fused; PKA, protein kinase A; GSK3, glycogen synthase kinase-3; CKI, casein kinase I; CBP, CREB binding protein; CiA, active form of Ci; CiR, repressor form of Ci.

1.9 The role of *GLI* genes in limb development

A multitude of studies in mice and other model organisms have proven that a GLI-code, the interplay of GLI proteins and the temporally fine tuned expression of the *GLI* genes in adjacent domains, together provide a basic tool that is used over and over again in embryonal development. To present a detailed insight into the early developmental events to which the members of *GLI* gene family participate is beyond the scope of this section; however roles of GLI3 in mammalian early limb patterning will be outlined.

The development of embryonic limb starts as a small bud-like structure containing a mass of mesenchyme cells covered by a layer of ectoderm. From the time once it appears, the limb bud undergoes gradual patterning along the

i) proximal-to-distal (P-D) axis, i.e. from shoulder to finger tip,

ii) anterior-to-posterior (A-P) axis, i.e. from thumb to little finger, ulna to radius in the forelimb, and

iii) dorsal-ventral (D-V) axis, i.e. back of the hand to palm.

This three-dimensional growth leads to the differentiation of the initially homogenous mass of mesenchyme cells into three morphologically distinct limb elements, i.e. stylopod (proximal element), zeugopod (intermediate elements), and autopod (the most distal elements).

During the early phase of limb development, the repressor form of Gli3 (GLI3R) is expressed throughout the bud, without A-P axis discrimination, which suppresses the expression of 5`*Hoxd* genes, *Shh* and *dHand* (Litingtung et al. 2002; te Welscher et al. 2002) (Fig. 1.18A). Subsequently, the basic-loop-helix transcription factor dHand restricts expression of *Gli3* within the anterior half of the emerging limb bud and is itself restricted to the posterior half by GLI3R. This dHand/GLI3 mutual antagonism is the basis of early molecular A-P asymmetry of the limb bud. In the posterior mesenchyme, dHand triggers the expression of the *Shh* and 5´*Hoxd* genes (Fig. 1.18B). dHand expression is localized to the posterior limb bud region before Shh and 5´Hoxd genes expression, and its inactivation abrogates expression of Shh (Charite et al. 2000; Fernandez-Teran et al. 2000; te Welscher et al. 2002) and posterior Hoxd genes (Yelon et al. 2000). The recently localized regulatory regions, i.e. ZRS (*zone of polarizing activity regulatory sequence*) (Lettice et al. 2003) and ELCR (*early limb control region*) (Zakany et al. 2004) have exclusively been associated with early limb specific expression of *Shh* and *Hoxd* genes (first wave of Hoxd expression), respectively, and dHand might interact directly or indirectly with these regulatory landscapes to position *Shh* and 5`*Hoxd* gene expression posteriorly (Fig. 1.18b). Conversely, the anterior restriction of these genes might be an effect of GLI3R on the dual nature ZRS (Sagai et al. 2005) and the exclusively repressor regulatory element POST (*posterior restriction,* Fig. 1.18B) (Tarchini and Duboule 2006) as in *Gli3*$^{-/-}$ mutant mice *Shh* and 5`*Hoxd* genes are ectopically expressed in the anterior limb mesenchyme (Sagai et al. 2004; te Welscher et al. 2002; Zuniga and Zeller 1999). Once activated within the posterior mesenchyme of limb bud, the 5`*Hoxd* genes (Hoxd13-10) activate the expression and precise localization of the Shh transcript possibly through the ZRS (Zakany et al. 2004). Thus, within the posterior region of the limb bud *Shh*, *dHand* and 5`*Hoxd* genes establish a dynamically interactive reciprocal

loop, in which 5`*Hoxd* genes activate Shh and dHand and the latter two activate each other as well as the posterior *Hoxd* genes (Fig 1.18C).

In addition, SHH and dHand impose constraints for GLI3R accumulation in the posterior region, either by inhibiting GLI3 to GLI3R conversion or directly repressing *Gli3* transcription (Fig. 1.18B & C) as in *dHand* and *Shh* mutants GLI3R accumulates over the whole limb bud (te Welscher et al. 2002). The early wave of *Hoxd* genes expression is sparked by two regulatory elements, one 3´ (ELCR) and other at the 5´ (POST) end of the cluster. The combinatorial affect of the two is critical to activate and restrict Shh expression within the posterior territory of the limb bud. This expression pattern is translated into proximal (stylopod) and intermediate elements (zeugopod) of the limb (Tarchini and Duboule 2006). Thus the early wave of Hoxd genes expression unifies the Phase-I and Phase-II of *Hox* gene expression as described above (Fig. 1.15).

Once triggered, SHH introduces a dramatic shift in the expression pattern of posterior *Hoxd* genes (*Hoxd9-13*) from the posteriorly restricted territory to a posterior-distal domain (Fig. 1.18C) which, constitutes the presumptive digit arch region. This SHH dependent second wave of 5´*Hoxd* genes expression is controlled by the GCR (*global control region*) (Spitz et al. 2003), positioned centromeric to the *Hoxd* cluster (Fig. 1.16 &1.19C). The second wave of *Hoxd* gene expression, which is triggered by SHH and controlled by the 5´located GCR leads to anterior-posterior asymmetry of the distal limb and the development of the autopod with digits. The second wave corresponds to Phase-III of *Hox* gene expression as described in the previous section (Fig. 1.15).

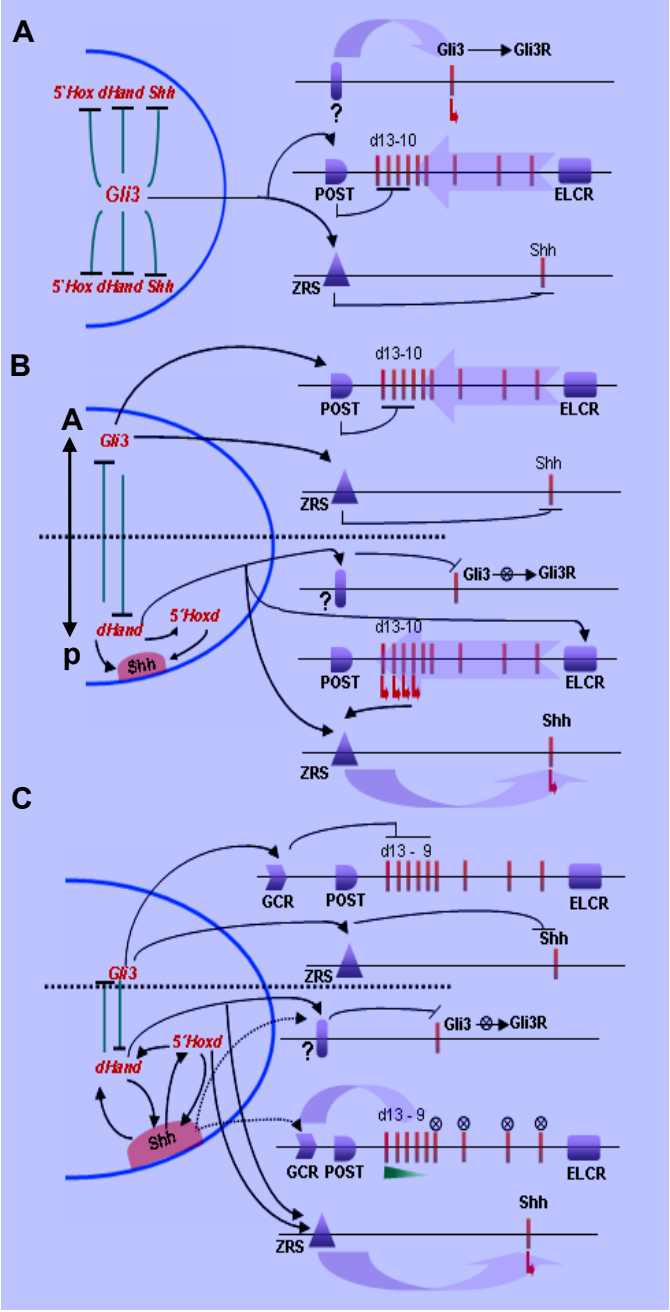


Figure. 1.18. Model for the direct or indirect interaction of Gli3, 5´Hoxd, dHand, and Shh with limb specific cis-acting regulatory elements during the establishment of early limb A-P asymmetry

Black arrows indicate the putative direct or indirect interaction of Gli3, 5´Hoxd, dHand, and Shh with the given cis-acting regulatory element. Black bars represent the negative influence on transcription while a broad lavender color arrow indicates the positive influence on transcription **A**) Gli3R is present throughout the limb bud, suppressing transcription of 5`*Hoxd* genes, *Shh* and *dHand*, **B**) reciprocal antagonism of Gli3 and dHand establish the early A-P asymmetry (dotted black line), dHand activates 5´*Hoxd* (*d13-10*) and *Shh* posteriorly; **C**) subsequently, in the posterior portion of the limb bud, positive feedback loops between 5´Hoxd, dHand, Shh trigger the progressive expansion of posterior identity. ELCR, early limb control region; POST, posterior restriction; ZRS, zone of polarizing activity (ZPA) regulatory sequence; GCR, global control region .

1.9.1 Gli3 in conjunction with Shh imposes in the autopod constraints on digit number and identity

GLI3 takes part in governing the development of all limb elements. It has been implicated in positioning the limb along the main body axis in combination with *dHand* and *Tbx3* (Rallis et al. 2005). Interaction of Gli3 and Alx4 as well as Plzf is essential for development of the

stylopod and the anterior elements of the zeugopod, i.e. femur, tibia and fibula (Barna et al. 2005; Panman et al. 2005).

Shh signaling acts in two ways to restrict Gli3R to the anterior domain of the developing limb bud, i) by preventing the formation of the truncated form of Gli3, and ii) by repressing its expression posteriorly. Thus sonic hedgehog creates a Gli3 gradient, such that the truncated form is seven times more abundant in the anterior third than in the posterior third of the emerging limb bud (Wang et al. 2000). In autopod patterning, this Shh-mediated anterior restriction of Gli3R is crucial for correct digit identity and number as evident from skeletal morphology of $Shh^{-/-}$, $Gli3^{-/-}$ and $Shh^{-/-}$: $Gli3^{-/-}$ null mutants (te Welscher et al. 2002) (Fig. 1.19).

In the $Shh^{-/-}$ mutants, there is a dramatic loss of intermediate and distal limb elements, with the zeugopod containing a single bone (radius/tibia) and a reduced digit 1 (Fig. 1.19). In the $Shh^{-/-}$ limb bud, Gli3R expression expands posteriorly, resulting in a down-regulation of posterior genes (*Gremlin*, *Fgf4*, 5`*Hoxd*) and initiating apoptosis, which leads to blockage of distal limb development and A-P patterning (te Welscher et al. 2002).

The limbs of $Gli3^{-/-}$ mutant mouse embryos show severe polydactyly with many unpatterned digits (Fig. 1.19), A-P polarity is lost and no apoptosis occurs in inter-digital regions (te Welscher et al. 2002). The polydactyly of Gli3 deficient limbs is Shh independent, as limbs lacking both Gli3 and Shh display a morphology (Fig. 1.19) identical to $Gli3^{-/-}$ mutants. Polydactyly in $Gli3^{-/-}$ and $Shh^{-/-}$: $Gli3^{-/-}$ double null mutants has been attributed to the anterior extension of Gremlin expression and general inhibition of BMP signaling. This contributes to the expansion of the mesenchyme and formation of additional digits (Aoto et al. 2002).

The abnormalities in limb patterning of the above mentioned null mutants suggest, that the default state of the limb is to form many digits. Gli3 and Shh impose the pentadactyly constraints so that five digits are formed in the mouse limb. Furthermore, Gli3 and Shh also specify the correct digit identity (Niswander 2003).

A study elucidating the cholesterol modification of Shh concluded that the formation of posterior digits 4 and 5 is governed by autocrine Shh signaling, while half of the digit 3 and the complete digit 2 is under the influence of paracrine Shh signaling (Lewis et al. 2001) (Fig. 1.19). In the case of autocrine signaling, the differential digit identities depend on the length of time the cells are exposed to high level of Shh signal. The identities of digits under paracrine signaling are concentration dependent. Digit 1 is not reliant on Shh, rather, a recent investigation suggests that digit 1 is specified by the interaction of the Alx4 transcription factor with Gli3 (Panman et al. 2005) (Fig. 1.19).

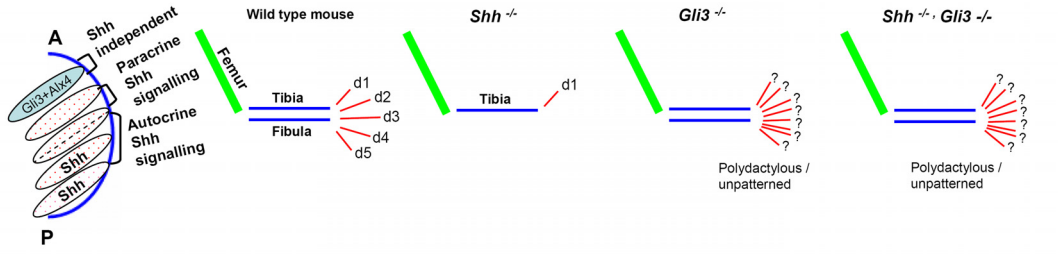


Figure. 1.19.The default state of the limb is to form many unpatterned digits, however, the genetic interactions between Shh and Gli3 define digit number and identity.

Schematic representation of limb elements of wild type and genetically manipulated mice. On the left side the developing limb bud is shown with red dotted ovals representing the presumptive digit-forming regions (5-4-3-2 from posterior to anterior direction) under the influence of Shh, while a blue oval depicts the presumptive digit1-forming region. Adopted and modified (Niswander 2003).

All modern tetrapods have limbs characterized by five or fewer digits (Tabin 1992). In contrast, recent fossil evidence indicates that Devonian tetrapods had greater number of digits. For example the *Acanthostega* forelimb had eight digits (Coates and Clack 1990). Thus, it appears that early tetrapods were polydactylous and the digit number has been reduced and stabilized at a maximum of five in the subsequent evolution of limb (Fig. 1.8) perhaps through the recruitment of Gli3 and Shh functions to impose constraints on autopod for digit numbers 5 or fewer.

The digit identity constraints, that is a maximum of five different types of digits might have been imposed earlier than the constraints on digit number itself, as *Acanthostega* possessed five morphological types of digits even though it had a total number of eight digits (Tabin 1992).

1.9.2 Gli3 functions other than limb morphogenesis

Gli3 functions are required for normal development of many organs and tissues, in particular the brain [dorsal telencephalon, diencephalon, midbrain, hindbrain; (Aoto et al. 2002)] but also the spinal cord (Litingtung and Chiang 2000) craniofacial structures, skeletal muscles (McDermott et al. 2005), lungs (Warburton and Lee 1999), eye (Aoto et al. 2002; Tyurina et al. 2005), and ear (Hui and Joyner 1993). It is also expressed in a wide variety of normal adult tissues, including heart, kidney, lungs, pancreas, liver, spleen, bone marrow, testis and myometrium (data from http://www.genecards).

1.10 The transcriptional regulation of the *GLI3* gene

GLI3 is an important developmental regulator and involved in the embryonic development of a long list of organs and also expressed in wide variety of adult tissues. The pleiotropy of GLI3 signals a highly complex and sophisticated cis-acting regulatory network to govern its expression in the correct spatiotemporal manner during development and postnatally. Recently, the minimal promoter of *GLI3* gene has been defined and three potential enhancer regions have been studied through a cell based reporter assay and by transient expression analysis in transgenic mouse embryos (Paparidis 2005). However, the majority of cis-acting regulatory elements (activators/repressors) which orchestrate the complex spatiotemporal expression of *GLI3* in juxtaposition with the promoter region have not been identified.

1.11 Paralogons, a mirror of chromosomal history or composed by functional restraints?

Paralogy regions (paralogons) are chromosomal regions each carrying a homologous set of genes belonging to different gene families. The human genome harbors several paralogons, with up to four copies of similar gene sets on different chromosomes, notably on HSA 1/6/9/19, HSA 4/5/8/10, HSA 1/2/8/10 and the HOX bearing chromosomes HSA 2/7/12/17. Relative order and transcriptional orientation of genes within paralogons can be maintained. Sometimes individual paralogous genes or groups of genes are scattered over the chromosome carrying the paralogon. Gene families found along the human paralogons may be entirely unrelated in both sequence and function, although occasionally there are hints of functional relatedness. Within a paralogon, some of the gene families may be represented on only one or two of the chromosomes, while others are represented on all four chromosomes. Sometimes further members of a gene family are found on other chromosomal regions outside the paralogons.

Along with other gene families, the distribution of the *GLI* gene family correlates with the chromosomal location of *HOX* gene clusters, with *GLI1*, *GLI2* and *GLI3* respectively, occurring close to the *HOXC*, *HOXD* and *HOXA* clusters (Fig. 1.20).

Primarily, there are two different hypotheses to explain the origin of the three or four fold paralogy seen on the present day human *HOX* cluster-bearing chromosomes and other paralogons. The first hypothesis suggests that they are the remnants of two rounds of duplication of chromosomal segments, whole chromosomes, or even the whole genome (2R hypothesis) (Larhammar et al. 2002).

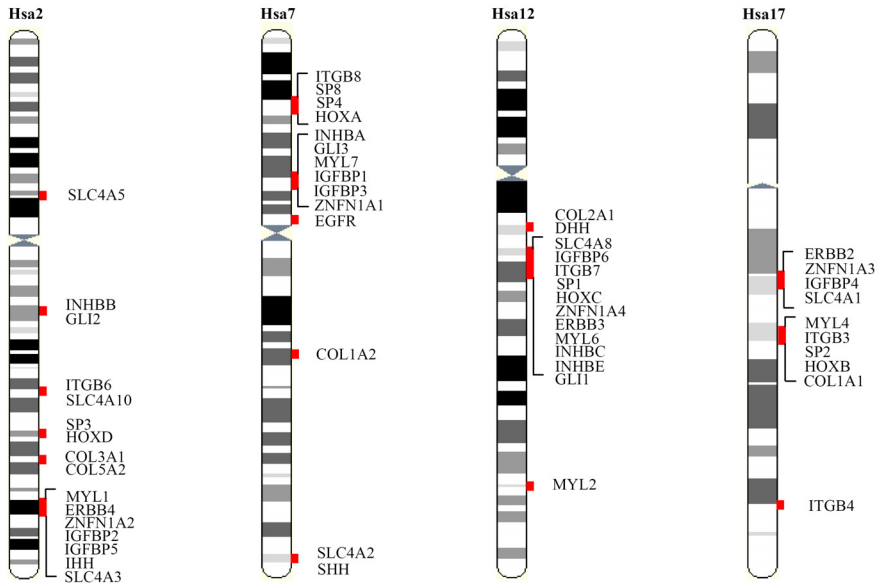


Figure. 1.20. Gene families including *GLI* with members on at least three of the human *HOX* bearing chromosomes 2, 7, 12, and 17

Restricted location of members of many of these gene familes near the *HOX* clusters suggests that these paralogons might have created by block duplication events. *SLC4*, solute carrier family 4; *INHB*, inhibins; *GLI*, glioma-associated oncogene homolog belonging to the krüppel family; *ITGB*, integrin β chains; *SP*, transcription factor Sp; *HOX*, homeo box; *COL*, collagens; *MYL*, myosin light chains; *EGFR/ERBB*, epidermal growth factor receptor/erythroblastoma; *ZNFN1A*, zinc finger protein, subfamily 1A; *IGFBP*, insulin-like growth factor-binding protein; *HH*, hedgehog. Features not drawn to scale.

The differences in the gene composition among chromosomal regions would be explained by gene loss events, while differences in gene order could be due to intrachromosomal inversions or translocations.

An alternative model, adoptive assembly, suggests that these paralogous gene sets have arisen largely by independent tandem or segmental duplications and translocation events (Hughes et al. 2001). In that case, the retention of collinear regions on different chromosomes is believed to reflect a selective advantage for relative positioning of certain genes (Furlong and Holland 2004).

1.12 Conservation of regulatory modules for paralogous genes?

The primary evidence that duplication of existing genes has played a vital role in the evolution of new gene functions is the widespread existence of gene families in modern vertebrates. Members of a gene family that share a common ancestor as a result of duplication event are denoted as being paralogous, distinguishing them from orthologous genes in

different genomes, which share a common ancestor as a result of speciation event. If genes duplicate in their entirety, i.e., each of the resultant copies include coding as well as all regulatory elements, then they can show inter-gene redundancy (Thomas 1993). The classical model predicts two different fates for newly duplicated functionally redundant duplicates: one copy accumulates degenerative mutations and thus becomes a functionally ineffective pseudogene (non-functionalization). Alternatively, one copy preserves the original ancestral function while the other copy by chance accumulates an advantageous mutation, conferring a novel positively selected function (neo-functionalization) (Force et al. 1999). According to this model (Fig. 1.21), the gene loss event is frequent, since only one of the duplicates is required to maintain the ancestral gene function, while the mutations that lead to novel functions are extremely rare. Thus the classical model predicts that few duplicates should be retained in the genomes over the long term. This model, however, fails to explain the existence of many duplicates in human and other vertebrate genomes (Li et al. 2001; Nadeau and Sankoff 1997).

A broadly accepted sub-functionalization model (Force et al. 1999) was proposed to explain the retention of the unexpectedly high number of duplicates in extant genomes: According to this second model, both duplicates undergo complementary degenerative changes, such that the duplicates together retain the original function of their ancestral gene (Fig. 1.21). This partitioning of ancestral functions among duplicates requires that the sub-functions are independent, in a way that the mutations in one will not affect another. These events might address independent subfunctions of an ancestral protein. However, the modular organization of eukaryotic enhancers makes them particularly suitable sub-functions. Furthermore the individual transcription factor binding sites are often short and degenerate, suggesting that they can frequently be degraded or created by simple point mutations. These properties of *cis*-acting regulatory sequences have led many researchers to hypothesize that sub-functionalization frequently acts upon regulatory sequences rather than by changing the coding contents. Thus a likely mechanism to retain multiple duplicates might involve the partitioning of ancestral expression domains of single gene among many descendents. During this process the duplicates independently evolve their ancient gene regulatory contents in a way that all of them together recapitulate the expression patterns of their ancestral gene (True and Carroll 2002).

One prominent example of subfunction partitioning via regulatory elements is mammalian EN1 and its zebrafish co-orthologs *eng1a* and *eng1b*. The fish co-orthologs are expressed in fin bud hindbrain, while mouse *En1* is expressed in both tissues. This suggests that most probably in the last common ancestor of fish and mammals this gene was expressed in both of

these tissues; however the fish specific duplication of this gene led to the selective loss of regulatory subfunctions among the duplicates (Postlethwait et al. 2004). Subfunction partitioning acting upon regulatory elements could entail regional (like *EN1*), quantitative, or temporal aspects of gene expression. In quantitative subfunction partitioning, degenerative mutations can decrease the amount of product so that the sum of expression of both genes is required to achieve the essential threshold amount of product. In temporal subfunction partitioning, both copies of ancestral gene can be expressed in the same tissue domain at different times by employing different subsets of regulatory elements (Postlethwait et al. 2004).

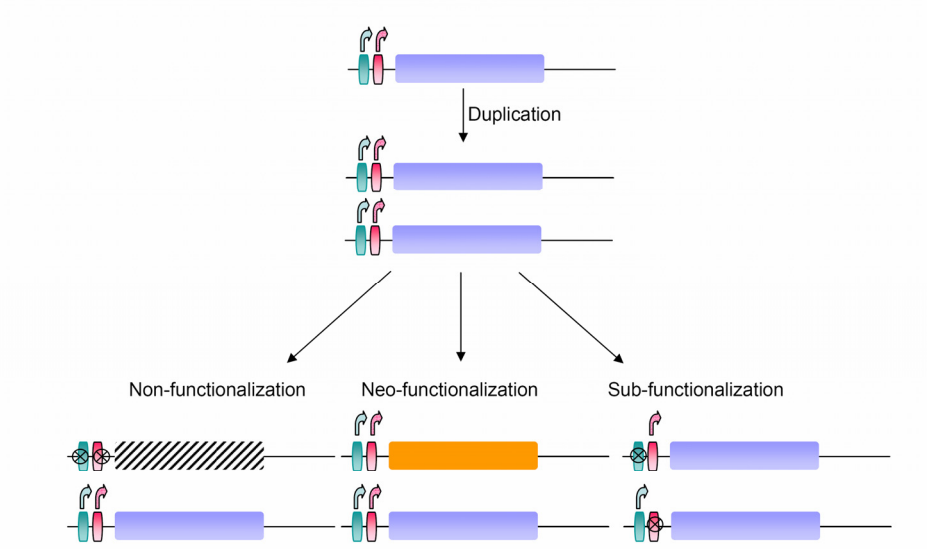


Figure. 1.21. Predicted fate of gene duplicates

The classical model predicts the loss of one copy through random accumulation of degrading mutations either in its coding or regulatory contents or gain of novel function through positive selection of advantageous mutations within coding contents.
The sub-functionalization model suggests that the likely mechanism to retain multiple duplicates involves the partitioning of ancestral expression domains of individual genes among many descendents through selective loss of regulatory subfunctions among the duplicates.

The GLI gene family members, i.e. the GLI1, GLI2 and GLI3 are known to have both unique and overlapping roles downstream of SHH in vertebrates. Furthermore, their expression domains are largely adjacent with GLI1 expression being preferably restricted to proliferating cells next to to Shh-expressing tissues, while GLI2 and GLI3 follow suite to GLI1 and are more broadly expressed in proliferating cells in regions more distant to Shh (Matise and Joyner 1999). Partial functional redundancy and the coordinated expression patterns suggest that in addition to constraints on coding intervals, the cis-acting regulatory

repertory might also remain largely preserved among three copies of ancestral gene following duplication events.

One example of transcriptional regulation by conserved enhancers can be found for the paralogs *Hoxa3*,and *Hoxb3*. Both genes contain in their enhancers binding sites for Krml-1, but in different copy numbers resulting in similarities as well as differences in the expression pattern of *Hoxa3* and *Hoxb3* in mouse hindbrain (Manzanares et al. 1999). *Hoxa4*, *Hoxb4*, and *Hoxd4* also share enhancer elements for mesodermal and neural regulation. These regulators are conserved also in the orthologs between human and mouse (Morrison et al. 1997).

The NDP kinase family similarly shows conservation of regulatory regions but with varying binding motifs, explaining the spatiotemporal differences in the expression of the different paralogs of this family (Ishikawa et al. 2003).

These examples show that paralogous genes can share common regulatory elements for overlapping expression. Still, differences in their placement and usage might allow distinct functions, as well.

1.13 Pathogenic effects of GLI mutations: GLI3 morphopathies

Changes in the expression of GLI genes due to regulatory defects or mutations affecting the function of the proteins are associated with pathogenic phenotypes both in man and mouse.

GLI1 is overexpressed in glioblastomas, osteosarcomas, rhabdomyosarcomas, B-cell lymphomas and basal cell carcinomas (Dahmane et al. 1997; Ghali et al. 1999; Kinzler et al. 1987; Kinzler and Vogelstein 1990; Roberts et al. 1989; Werner et al. 1997). Ectopic expression of human *GLI1* in mice has been observed to cause developmental defects, failure to thrive and Hirschsprung-like dilatation of the gastrointestinal tract (Yang et al. 1997).

Defects in GLI2 in mouse can cause basal cell carcinomas and skeletal disorders (Grachtchouk et al. 2000; Park et al. 2000; Sasaki et al. 1999). In humans, loss-of-function mutations can lead to pituitary anomalies and holoprosencephaly-like features (Roessler et al. 2003).

Point mutations, translocations and deletions throughout and flanking the *GLI3* gene can cause various autosomal dominant polysyndactyly syndromes such as the Greig cephalopolysyndactyly syndrome (GCPS), which mainly features preaxial polydactyly, syndactyly, broad thumbs and toes facial deformities such as hypertelorism and frontal bossing (Kalff-Suske et al. 1999; Vortkamp et al. 1991; Wild et al. 1997). Frameshift and nonsense mutations can also lead to Pallister-Hall syndrome (PHS) which is characterized by

hypothalamic hamartoma, central or postaxial polydactyly, syndactyly, imperforate anus, anteverted nares and other facial abnormalities and occasionally, associated HPE and malformations of the axial skeleton (Kang et al. 1997a; Kang et al. 1997b). Finally, nonsense and missense *GLI3* mutations can cause postaxial polydactyly type A, preaxial polydactyly type IV, and postaxial polydactyly type A/B. Until lately, it was supposed that there was no correlation between the site and the type of the mutation and the phenotype (Kalff-Suske et al. 1999), but recently it has been postulated that the site of frameshift and nonsense mutations can actually play a role in the determination of the syndrome. If the stop-mutation happens in the first third of the *GLI3* gene it may cause GCPS, whereas mutations in the second third are associated with PHS. Some GCPS patients have been found to be mutated in the third part of the gene, however, no PHS patients are associated with changes in this section (Johnston et al. 2005).

1.14 AIMS

Human GLI3, a member of the GLI family of transcription factors, is a key developmental regulator and dynamically expressed in brain, axial, appendicular, and craniofacial structures, as well as within numbers of visceral organs prenatally, postnatally and in adult life. The complex spatiotemporal and quantitative aspects of *GLI3* 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.

In this study, in an attempt to define and characterize the cis-acting regulatory repertory of *GLI3*,

i) as candidate enhancers for *GLI3* expression control, tetrapod-teleost conserved non-coding elements (CNEs) associated with *GLI3* will be localized through multi-species comparative sequence analysis.

ii) The potentially *GLI3*-specific gene regulatory role of these anciently constraint intervals will be tested by reporter gene expression based *in vitro* and *in vivo* experiments, employing transfected cell cultures as well as transgenic embryos of the model organisms zebrafish, chick and mouse.

iii) Functionally relevant *cis*-acting elements will be scrutinized for human/*Fugu* conserved transcription factor binding sites (TFBSs) by combining the phylogenetic footprinting with pattern recognition programs.

iv) To zoom in on functionally essential modules within a subset of *GLI3*-associated enhancers, deletion constructs will be analyzed in transfected cell cultures for their regulatory potential on reporter genes.

v) *GLI* family members are paralogs with partly overlapping functions and highly related expression patterns, and they are located in paralogons associated with the *HOX* clusters. To estimate the extent of functional constraints operating on members of the *GLI* gene family in vertebrates following the duplication events, the *GLI* gene regions will be compared to search for hints at retention of regulatory elements together with the coding sequences.

vi) To address the evolutionary events which shaped the four-fold paralogy regions in the human HOX-bearing chromosomes (HOX cluster paralogon), a phylogenetic analysis of those multigene families (including the *GLI* paralogs) will be performed having members on at least three of the human HOX clusters bearing chromosomes (HSA 2/7/12/17) (test of 2R hypothesis versus small scale duplication hypothesis).

MATERIALS & METHODS

2.1 Reporter constructs

Candidate enhancer sequences (CNEs) were PCR amplified (Table 2.1) using the high fidelity herculase enhanced DNA polymerase (Stratagene, USA) with primers containing KpnI restriction site tags. Amplified DNA was purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany). Purified PCR products were then subjected to restriction site digestion with KpnI (New England Bio Labs, Ipswich, USA) and subsequently cloned in both orientations into reporter constructs upstream of a minimal *GLI3* promoter or a heterologous SV40 promoter driving expression of the firefly luciferase gene in the vector pGL3 (Promega, Madison, USA). The reporter constructs were designated pGL3-promGLI3-300-luc and pGL3-promSV40-luc, respectively.

To generate transgenic mice, the PCR-products PvuII_GLI3 (promoter proximal region, NCBI Build 36.1, coordinates Chr7: 42243912-42247420) and CNE1, 6, 9, 10, 11 (Table 2.1) were cloned into the multiple cloning site (CNE1, 6, 9 and 10 at the Kpn1 site, CNE11 at the Pst1 site and PVUII_GLI3 at the Kpn1-Xho1 sites) of the p1230 vector (Lettice et al. 2003) in front of the human β-globin promoter driving an X-gal reporter gene. In transgenic mice, the β-globin promoter does not drive expression of the reporter gene by itself (Simmons et al. 2001). Recombinant reporter expression constructs were transfected into Top10 competent bacterial cells (Invitrogen, Karlsruhe, Germany) and subsequently isolated and purified using the Qiagen plasmid purification kit (Qiagen, Hilden, Germany). To control the clones for presence of any point mutations generated during the PCR amplification, appropriate DNA preparations were sequenced in an ABI 377 automated sequencer (Applied Biosystems, Foster City, California, USA) and were analyzed with Sequencher, Version 4.2, software (Gene Codes, Ann Arbor, Michigan, USA).

2.2 Deletion mutants

The deletion mutants of selected CNEs were made by PCR (Table 2.2) using the recombinant reporter constructs of each of the respective wild type CNE as a template. The sequences flanking the segment to be deleted were PCR amplified with two different sets of primers. One member of each set was WT primer tagged with a KpnI restriction site, while the other member was designed from immediate vicinity of the sequence to be deleted and tagged with a HindIII restriction site. Amplified products flanking the region to be deleted were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany), digested by HindIII, and subsequently ligated (at HindIII digested ends). The ligated products were

size fractioned on 2% agarose gel, and the DNA fragment of expected length was gel excised, purified by using a QIAquick gel extraction kit (Hilden, Germany), subsequently digested by KpnI, and inserted into the pGL3-promGLI3-300-luc reporter plasmid. The DNA sequence of each deleted recombinant construct was confirmed. In order to avoid the de-novo creation of transcription factor binding sites, compared to wild type sequence, each of the prospective deleted sequences was analyzed before cloning for potential TFBS with the TESS web tool (Transcription element search software on http://www.cbil.upenn.edu/tess).

Table 2.1. Primers used to amplify the intra-*GLI3* conserved non-coding elements (CNEs) and PvuII_GLI3

Element	Forward primer	Reverse primer	Annealing temperature	Restriction site Tag
CNE1	5-gcggtacccttaggagaccattcccacatgg-3	5-gcggtaccctccctcatcaggtgatcaatg-3	65 °C	Kpn1
CNE2	5-cgaagcttgagcaattgcagagtgcagg-3	5-gcaagcttcacctctctccaaccagctag-3	74 °C	Hind-III
CNE3	5-gcaagcttgaatgtcgcagggcagaaaatg-3	5-gcaagcttcatggtgccttgtacaggtg-3	72 °C	Hind-III
CNE4	5-gcaagcttcatcaacgatatggtgcag-3	5-gcaagcttgtctgtaattgcagttgttc-3	69 °C	Hind-III
CNE5	5-gcggtacccgagcacttcttcagcttggaac-3	5-gcggtacccagtactgagccacaagcttcaac-3	65 °C	Kpn1
CNE6	5-gcggtaccctcgtgcctctgagcagaaagg-3	5-gcggtaccgtcttcctggtggcctttcc-3	65 °C	Kpn1
CNE7	5-gcggtaccctacagccttagtctgcttacc-3	5-gcggtaccgtgccttaagaactttccagc-3	65 °C	Kpn1
CNE9	5-gcggtacccacatttccacagcagatattatg-3	5-gcggtacccacaaatgcactcacaattagg-3	58 °C	Kpn1
CNE10	5-gcggtaccgttgatccccaaatactggattc-3	5-gcggtacccaatcagcagaaatctactaaatccc-3	60 °C	Kpn1
CNE11	5- gcacgcgtctggacttctcaaaatgtcagg-3	5- gcacgcgtcacttcatacagtggatggagag-3	65 °C	Mlu1
CNE12	5-gcggtaccggtttcaaagcaaataaagggagacag-3	5-gcggtacccaatttctttcaaacagctctcc-3	70 °C	Kpn1
PvuII_GLI3	5-gcggtaccctgggaatgcaaactctcagattg-3	5-gcctcgaggctttctcgctccttccagg-3	72 °C	Kpn1-Xho1

*Note:*CNE8 primers are not given; despite of extensive efforts this interval was not amplified from genomic DNA because of unknown reasons.

Table 2.2. Primers used to generate the deletion constructs for selected CNEs

Element/Primers	Sequence	Annealing temperature	Restriction site tag
CNE1			
up-UC-CNE1-Fwd	5-gcggtacccttaggagaccattcccacatgg-3	65°C	Kpn1
up-UC-CNE1-Rev	5-gcaagcttgctacaaagatggctggggac-3		HindIII
down-UC-CNE1-Fwd	5-gcaagcttggaaaatattgaatggactcagagc-3	65°C	HindIII
down-UC-CNE1-Rev	5-gcggtaccctccctcatcaggtgatcaatg-3		Kpn1
CNE1-UC-Fwd	5-gcggtaccctgataaggtttcatttcagaatg-3	58°C	Kpn1
CNE1-UC-Rev	5-gcggtaccctgagtccattcaatattttccac-3		Kpn1
CNE5			
up-UC-CNE5-Fwd	5-gcggtacccgagcacttcttcagcttggaac-3	63°C	Kpn1
up-UC-CNE5-Rev	5-gcacgcgtccctctcagagcctgtgctatg-3		Mlu1
down-UC-CNE5-Fwd	5-gcacgcgtcagtttagcgattttatcaggcac-3	63°C	Mlu1
down-UC-CNE5-Rev	5-gcggtacccagtactgagccacaagcttcaac-3		Kpn1
CNE5-UC-Fwd	5-gcggtaccgggtaagacttcatgattaatgg-3	63°C	Kpn1
CNE5-UC-Rev	5-gcggtaccgtgactctcctcaatttacactgc-3		Kpn1
CNE6			
up-UC-CNE6-Fwd	5-gcggtaccctcgtgcctctgagcagaaagg-3	60°C	Kpn1
up-UC-CNE6-Rev	5-gcacgcgtgattgtgacgtgatccataatcac-3		Mlu1
down-UC-CNE6-Fwd	5-gcacgcgtgaatccaccttgctcctctcc-3	60°C	Mlu1
down-UC-CNE6-Rev	5-gcggtaccgtcttcctggtggcctttcc-3		Kpn1
CNE6-UC-Fwd	5-gcggtaccgattctaaaatccatcttggtgg-3	60°C	Kpn1
CNE6-UC-Rev	5-gcggtaccgatcacgtcacaatctaatagcc-3		Kpn1

2.3 Cell cultures

The human lung tumor cell line H661 and the human bronchiolar epithelial cells H441 were obtained from the ATCC (American Type Culture Collection), USA, and grown at standard conditions in RPM1-1640 medium (Sigma Aldrich, Missouri, USA) containing 10% fetal calf serum, 1% non-essential amino acids, 2% penicillin / streptomycin, and 1% L-glutamine (H661), or in modified RPMI-1640 medium (Sigma Aldrich, Missouri, USA) with 25mM HEPES and sodium bicarbonate, containing 4% fetal calf serum, 1% non-essential amino acids, 2% penicillin / streptomycin and 1% L-glutamine (H441), respectively.

2.4 Transient transfection and dual luciferase assay

The day before transfection, 4×10^5 H661 or 3×10^5 H441 cells were seeded into each well of a 12-well plate (Greiner Bio-One, Frickenhausen, Germany) in 2 ml of the appropriate growth medium containing serum and antibiotics. After 24 hours of incubation at standard growth conditions, cells were transfected by using Effectene (Qiagen, Hilden, Germany) according to the manufacturer's recommendations with the pGL3-derived experimental firefly luciferase reporter constructs at a concentration of 200ng/well, along with 100ng/well of pRLSV40 (Promega, Madison, USA), an expression vector containing cDNA encoding *Renilla* luciferase as an internal control reporter, and 200ng/well of pGKBT7 (Clontech, California, USA) as a stuffer/carrier DNA.

48 hours after transfection, cells were assayed for luciferase activity with the Dual-Luciferase Reporter Assay System (Promega, Madison, USA) on an AutoLumat LB 953 luminometer (Berthold, Germany). The activities of experimental reporter (firefly luciferase) were normalized to the activities of internal control reporter (*Renilla* luciferase). All assays were conducted three times in triplicate.

2.5 Zebrafish enhancer / GFP reporter Assay

Zebrafish were bred and raised according to standard protocols (Westerfield 2000). CNEs for co-injection were either cut out from plasmids or amplified by PCR, and then purified by the QIAquick PCR purification kit (Qiagen, Hilden, Germany). The reporter expression construct consisting of cDNA encoding Enhanced Green Fluorescent Protein (EGFP) under the control of a minimal promoter from the mouse β-globin gene was also PCR amplified from a plasmid construct (Clontech; Palo Alto, California, USA). Element DNA (250-300ng/µl) and reporter DNA fragment (25ng/µl) were combined with tracer, i.e. phenol red (0.1%), and co-injected with a femtojet pressure injection system (Eppendorf; Hamburg, Germany) into the embryos at the 1- to 8-cell stage produced from natural mating essentially as described (Woolfe et al. 2005). Embryos developing abnormally were discarded after 2 to 3 hours of injection. Normal embryos were raised in 0.003% phenylthiocarbamide from tailbud stage. On the second day of microinjection (approximately 26-33 hpf), embryos were dechorionated using pronase E, anaesthetized in Tricaine, and analysed under UV-light for GFP expression by using an Olympus IX81 motorised inverted microscope (Tokyo, Japan). Images were captured using an FVII CCD monochrome digital camera and analySIS image-processing software (Olympus, Tokyo, Japan).

GFP expressing cells were classified according to the following tissue categories: forebrain, midbrain, hindbrain, spinal cord, eye, ear, notochord, muscle, blood (circulating) /

blood islands, heart / pericardial region, epidermis, and fins. GFP expressing cells that were not localized unequivocally were classified as "others". Location and tissue category of each GFP-expressing cell for each embryo was recorded schematically using Adobe Photoshop software (Adobe Systems, San Jose, California, USA), onto an overlay of a camera lucida drawing of a 31-hpf embryo. For each CNE, the GFP expression data was collected from 20-50 expressing embryos. As a control, mean of 200 embryos were injected with conserved coding and non-conserved intronic sequences along with the reporter system and were found unable to show any significant GFP induction (Woolfe et al. 2005).

Combined schematised expression data for each CNE was compressed into a JPEG file and coupled with graphical depiction of expression domains to present an overall impression of the spatial pattern to which the element directs expression.

2.5.1 Anti-GFP Immunostaining

For immunostaining, embryos were fixed in 4% paraformaldehyde overnight at 4°C and incubated with rabbit polyclonal anti-GFP antibody (AMS Biotechnology, Abingdon Oxon, UK) using standard protocols (Moens and Fritz 1999) and the ABC amplification system (Vectastain; Vector laboratories, Burlingame, California, USA). Stained embryos were subsequently cleared in glycerol, flatmounted, and observed under bright field with an Olympus IX81 motorised inverted microscope (Olympus, Tokyo, Japan).

2.6 Chicken in ovo electroporations and enhancer reporter expression analysis

1μg/μl of each GFP reporter construct (Bg-EGFP) containing the conserved non-coding element was co-electroporated together with 1μg/μl of an RFP expression vector [RFP in pCAGGs driven by the U6 promoter from chick chromosome 28 (Das et al. 2006) and 0.02% fast green. This mix was injected into the limb bud mesenchyme at stages HH 19/20 and electroporated with 1 pulse of a square wave current generated by a CUY21 electroporator (Bex, Tokyo, Japan) at 45 volts for 50 msec using 3mm platinum electrodes placed anterior and posterior to the limb bud. Electroporation into the presumptive limb mesenchyme was carried out by injecting the RFP reporter and putative enhancer constructs into the coelom at stage HH14. The positive electrode was placed lateral to the coelom and the negative electrode above the neural tube and a square wave current of 60 volts for 50 msec with a 500 msec pause was applied 5 times. Limb buds were analyzed as whole mounts for GFP and RFP expression 48 hours following electroporation using an UV fluorescence dissecting microscope and a GFP or TXR filter, respectively, when the embryos had reached

approximately stage HH 26 after electroporation at HH19/20 or HH23 after electroporation at HH14.

2.6.1 In situ hybridization

Chick embryos were fixed in 4% paraformaldehyde at 4°C and then dehydrated in 100% methanol. Whole mount *in situ* hybridization was performed as described (Riddle et al. 1993).

2.7 Generation of transgenic mice

50-100 μg of plasmid was double digested at the sites flanking the CNE-lacZ-SV40 system to slice out this system from remaining part of recombinant p1230 vector. The digestion product of each reaction was separated in an 1% agarose gel, and the DNA band corresponding to the insert was extracted with the QIAquick gel extraction method. The isolated, linearized DNA fragment was then precipitated with 1/10 volume of sodium acetate and a double volume of 100% EtOH, and incubated for 30' at -80^{0}C. The mixture was centrifuged at 14.000 rpm at rome temperature for 15'. The supernatant was removed and the pellet was washed with 200 μl of 70% EtOH. The isolated DNA fragments were supplied for microinjection diluted in 10 mM Tris, pH 7.5, 0.1 mM EDTA, pH 8.0 buffer in a final concentration of 1-3 μg/ml.

Transgenic mice were generated by commercial suppliers. The donor eggs originated from FVB females (approx 3 weeks old and superovulated with hormones), which were crossed with FVB males. The eggs were transferred into the oviduct of CBA females (crossed with sterile CBA males). The amount of DNA injected cannot be determined with certainty, but it is estimated that 1-2 pl are microinjected into each male pronucleus of fertilized eggs. Biopsies of the offspring were supplied to be tested by PCR for the presence of the transgene as described (Paparidis 2005). Positive (transgenic) mice were maintained in the animal facility of the "Fachbereich Medizin der Philipps Universität Marburg". Male transgenic mice were crossed with wild-type females to obtain embryos.

2.7.1 Embryo staining and histological analysis

The time of gestation was calculated taking noon of the day of detection of a vaginal plug as embryonic day 0.5 (E 0.5). Embryos were harvested at E9.5, 10.5, 11.5, 12.5, and 13.5. Expression of the transgene was detected by staining staged embryos overnight in X-gal buffer. They were dissected free of extraembryonic membranes (which were retained for genotype analysis) then fixed in 0.5% glutaraldehyde at 4^{0}C for 30' to 2 hours, depending on their developmental stage, washed with PBS (containing 2mM $MgCl_2$), and stained overnight in X-gal reaction buffer [(containing 35 mM $K_3Fe(CN)_6$, 35 mM $K_4Fe(CN)_6$ and 2mM

MgCl2)] containing 0.1% Xgal at 37^0C. The staining reaction was stopped by washing the embryos in PBS. The embryos were postfixed overnight in 0.5% glutaraldehyde at 4^0C.

To analyze internal organs, some of the embryos were dehydrated, embedded in paraffin wax, and sectioned at 10-40 µm following standard protocols. Sections were deparaffinized with xylene and mounted for histological analysis.

Genomic DNA (extracted from tail or ear tissue using standard protocols) was used, for genotyping of the mice carrying recombinant constructs by PCR, employing the primers "XgalF",5'-CAACAGTTGCGCAGCCTGAATG-3'," XgalR", 5'-GTGGGAACAAACGGCGGATTG -3' (Paparidis 2005).

2.8 Sequence data and comparative analysis

Approximately 1Mb of the human genome, encompassing *GLI3* (ENSG00000106571) as well as *GLI3* orthologous sequences of Chimpanzee (ENSPTRG00000019117), Rhesus (ENSMMUG00000013614), Mouse (ENSMUSG00000021318), Rat (ENSRNOG00000014395), Dog (ENSCAFG00000003535), Cow (ENSBTAG00000010671) Chick (ENSGALG00000012329), Opossum (ENSMODG00000002714), Frog (ENSXETG00000001856), and *Fugu* (SINFRUG00000153715) were retrieved from the ENSEMBL genome browser (http://www.ensembl.org).

Multispecies sequence comparison was performed by using the Shuffle-LAGAN alignment tool kit (Brudno et al. 2003a). Human sequence was used as the baseline and annotated by using the exon / intron information available at ENSEMBL genome browser. Shuffle-LAGAN alignment was visualised with the VISTA visualization tool (Mayor et al. 2000). The conservation was measured using a 60bp window and a cutoff score 50% identity, with human sequence as a base line.

2.9 In silico mapping of conserved transcription factor binding sites within each CNE

Human-*Fugu* conserved transcription factor binding sites in each CNE were detected with ConSite (http:/www.phylofoot.org/consite) and rVISTA.2.0 (http:/rvista.decode.org/). The ConSite screen for conserved TFBS was performed against the JASPAR database (http://jaspar.genereg.net) with 60% conservation cutoff, 50 base-pair window size, and 75% transcription factor score threshold settings.

rVISTA 2.0 searches for conserved TFBSs were performed against 500 vertebrate TF matrices from the TRANSFAC library (Matys et al. 2006), with a matrix similarity cut-off at 0.85, by submitting the BlastZ alignment file for each CNE to the rVISTA 2.0 site.

2.10 Estimation of evolutionary constraints on *GLI* sequences in vertebrates

In order to analyze the evolutionary relationship of *GLI* sequences between and within species, the complete cDNAs and corresponding protein sequences (Table 2.3) for human *GLI* gene family members, i.e. *GLI1*, *GLI2,* and *GLI3,* and their orthologs in mouse, rat, *Fugu*, tetraodon, zebrafish, D. melanogaster were extracted from ENSEMBLE genome browser (http://www.ensembl.org).

Table 2.3. ENSEMBL derived peptides and cDNAs used to analyze the sequence evolutionary patterns of *GLI* genes

Sequence	Peptide ID	Transcript ID
Human		
GLI1	ENSP00000228682	ENST00000228682
GLI2	ENSP00000354586	ENST00000361492
GLI3	ENSP00000265526	ENST00000265526
Mouse		
GLI1	ENSMUSP00000026474	ENSMUST00000026474
GLI2	ENSMUSP00000070591	ENSMUST00000063361
GLI3	ENSMUSP00000021754	ENSMUST00000021754
Rat		
GLI1	ENSRNOP00000009803	ENSRNOT00000009803
GLI2	ENSRNOP00000009963	ENSRNOT00000009963
GLI3	ENSRNOP00000019396	ENSRNOT00000019396
Fugu		
GLI1	NEWSINFRUG00000154410	NEWSINFRUT00000164302
GLI2	NEWSINFRUP00000159280	NEWSINFRUT00000159280
GLI3	NEWSINFRUP00000163565	NEWSINFRUT00000163565
Tetraodon		
GLI1	GSTENP00013570001	GSTENT00013570001
GLI2	GSTENP00033101001	GSTENT00033101001
GLI3	GSTENP00025555001	GSTENT00025555001
Drosophila		
Ci	CG2125-PA	CG2125-RA

The numbers of synonymous nucleotide substitutions per synonymous (Ks) and non-synonymous nucleotide substitutions per non-synonymous site (Ka) were calculated by using the Li-Wu-Lu method (Li et al. 1985) in pairwise comparison.

Evolutionary distance between all possible pairs of GLI paralogs within each lineage was estimated by Tajima's relative rate test (Tajima 1993).

2.11 Dataset for gene families linked with the human HOX clusters

Genes from 11 families were included in the analysis (Table 2.4). The chromosomal location of human gene families was obtained from the ENSEMBL genome browser. Eight of these families have members on each of the human HOX-bearing chromosomes while 3 have their members on at least three of those chromosomes (Table 2.4). Information about the molecular functions of selected gene families (Table 2.4) was retrieved from GeneReports available at SOURCE (http://source.stanford.edu).

The closest putative orthologous sequences of human proteins in other species were obtained from Orthologue Prediction at ENSEMBL. To enrich these gene families with sequences from those organisms for which the sequence information was not available at ENSEMBL, a BLASTP search (Altschul et al. 1990) was carried out against the protein database available at National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov) and the Joint Genome Institute (http://www.jgi.doe.gov). Because the focus of this study was to identify the duplication events which had occurred during vertebrate evolution, only blast hits giving a higher score than the sequence of available invertebrate ancestral sequences were retained. Further confirmation of ancestral-descendents relationship among putative orthologs was done through clustering of homologous proteins within phylogenetic trees. I excluded sequences whose position within a tree was sharply in conflict with the uncontested animal phylogeny. The list of all used sequences is available online. (http://www.biomedcentral.com/imedia/9632252721640605 /supp1/.doc).

The species we chose are *Homo sapiens (human)*, *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Gallus gallus* (chicken), *Macaca mulatta* (rhesus monkey), *Canis familiaris* (dog), *Bos taurus* (cow), *Monodelphis domestica* (opossum), *Xenopus tropicalis* (Frog), *Erinaceus europaeus* (hedgehog), *Danio rerio* (zebrafish), *Takifugu rubripes* (Fugu), *Tetraodon nigroviridis*, *Gasterosteus aculeatus* (Stickleback), *Oryzias latipes* (Medaka), *Ciona intestinalis* (ascidian), *Ciona savignyi* (ascidian), *Branchiostoma floridae* (Amphioxus), *Strongylocentrotus purpuratus* (sea urchin), *Drosophila melanogaster* (fruit fly), *Apis mellifera* (honey bee), *Caenorhabditis elegans* (Nematode).

Table 2.4. Human gene families used in analysis

Gene family	Members	Chromosome location	Human protein accession No	Molecular function
Fbrillar collagen family				
	COL2A1	12q13.11-q13.2	P02458	Extracellular matrix structural constituent, Structural constituent of bone, Phosphate transport, Cell adhesion,Skeletal development, Perception of sound.
	COL3A1	2q31	P02461	
	COL5A2	2q14-q32	Q7KZ55	
	COL1A2	7q22.1	P08123	
	COL1A1	17q21.33	P02452	
ERBB receptor protein-tyrosine kinase				
	ERBB2	17q21.1	Q96RT1	Epidermal growth factor receptor activity, Protein serine/threonine kinase activity, Electron transporter activity, Cell proliferation, ATP binding.
	EGFR	7p12	P00533	
	ERBB4	2q33.3-q34	Q15303	
	ERBB3	12q13	P21860	
Insulin-like growth factor-binding protein				
	IGFBP4	17q12-q21.1	P22692	Regulation of cell growth, Signal transduction, Skeletal development Cell proliferation.
	IGFBP1	7p13-p12	P08833	
	IGFBP2	2q33-q34	P18065	
	IGFBP6	12q13	P24592	
	IGFBP3	7p13-p12	P17936	
	IGFBP5	2q33-q36	P24593	
Integrin beta chain family				
	ITGB3	17q21.32	P05106	Receptor activity, Cell-matrix adhesion, Integrin-mediated Signaling pathway.
	ITGB5	3q21.2	P18084	
	ITGB6	2q24.2	P18564	
	ITGB7	12q13.13	P26010	
	ITGB4	17q25	P16144	
	ITGB8	7p15.3	P26012	
Myosin light chain				
	MYL4	17q21-qter	P12829	Phosphoprotein phosphatase activity, Structural constituent of muscle, Muscle development, Microfilament motor activity.
	MYL6	12q13.2	P60660	
	MYL1	2q33-q34	P06741	
	MYL7	7p21-p11.2	Q01449	
	MYL2	12q23-q24.3	P10916	

Table 2.4 Continued

Sp1 c2h2-type zinc-finger protein family				
	SP1	12q13.1	P08047	RNA polymerase II transcription factor activity.
	SP2	17q21.32	Q02086	
	SP3	2q31	Q02447	
	SP4	7p15	Q02446	
	SP8	7p21.2	Q8IXZ3	
Zinc finger protein, subfamily 1A				
	ZNFN1A1	7p13-p11.1	Q13422	DNA-dependent regulation of transcription, Specification and the maturation of the lymphocyte.
	ZNFN1A2	2qter	Q9UKS7	
	ZNFN1A3	17q21	Q9UKT9	
	ZNFN1A4	12q13	Q96JP3	
Anion exchanger family SLC4A (AE)				
	SLC4A1	17q21-q22	P02730	Inorganic anion exchanger activity, Bicarbonate transport, Chloride transport.
	SLC4A2	7q35-q36	P04920	
	SLC4A3	2q36	P48751	
	SLC4A5	2p13	Q14203	
	SLC4A8	12q13	O95233	
	SLC4A10	2q23-q24	Q9HCQ6	
GLI zinc-finger protein family				
	GLI1	12q13.2-q13.3	P08151	Regulation of transcription from RNA polymerase II promoter, Morphogenesis of limb and brain.
	GLI2	2q14	P10070	
	GLI3	7p13	P10071	
Hedgehog family				
	SHH	7q36	Q15465	Mesodermal cell fate determination, Proteolysis and peptidolysis, Cell-cell signaling, Intein-mediated protein splicing.
	DHH	12q12-q13.1	O43323	
	IHH	2q33-q35	Q14623	
Inhibin				
	INHBA	7p15-p13	P08476	Cytokine activity, Growth factor activity, Induction of apoptosis, Mesoderm development, Defense response.
	INHBB	2cen-q13	P09529	
	INHBC	12q13.1	P55103	
	INHBE	12q13.3	P58166	

2.12 Alignment and phylogenetic analysis of gene families linked with the human HOX clusters

Amino acid sequences were aligned by using CLUSTAL W (Thompson et al. 1994) under default parameters. The alignments were manually refined where necessary. The phylogenetic trees for each gene family were reconstructed by using the neighbor-joining (NJ) method (Saitou and Nei 1987), the complete deletion option was used to exclude any site which postulated a gap in the sequences. Poisson corrected (PC) amino acid distance and uncorrected proportion (*p*) of amino acid difference were used as amino acid substitution models. Because both methods produced similar results, only the results from NJ tree based on uncorrected *p*-distance are presented here. Reliability of the resulting tree topology was tested by the bootstrap method (Felsenstein 1985) (at 1000 pseudoreplicates) which generated the bootstrap probability for each interior branch in the tree.

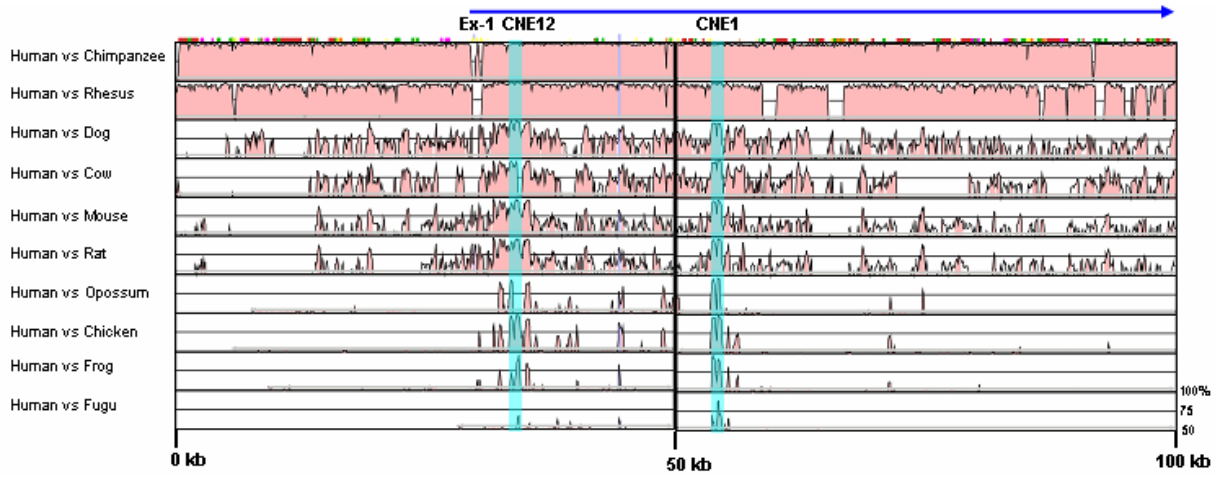
The phylogenetic trees of seven gene families (COL, ERBB, IGFBP, ZNFN1A, GLI, HH and INHB) were rooted with orthologous genes from invertebrates, whereas the SP phylogeny was rooted with both invertebrate and vertebrate SP8 sequences. The phylogenies of SLC4A and MYL families consisted of two subfamilies, each of which served to root the other. For the ITGB tree, vertebrate ITGB8 sequences served as an outgroup to root the remainder of the tree, while the remaining sequences served to root vertebrate ITGB8 sequences.

For each gene family the order of branching within the phylogenetic tree was used to estimate the time window for gene duplication events relative to the divergence of major taxa of organisms. This method of relative dating does not depend on the assumption of a constant rate of molecular evolution and is thus robust to differences in the rate of evolution in different branches of the tree (Hughes 1998). Tree topology of each gene family was compared with other families and also with the phylogeny of HOX clusters (Zhang and Nei 1996) to test consistencies in duplication events.

RESULTS

3.1 Prioritization of intra-*GLI3* CNEs (conserved non-coding elements) for functional analysis

In the pufferfish, *gli3* (scaffold_210; ENSEMBL genome browser) is tightly bordered by genes not orthologous to the human *GLI3* flanking regions. Therefore, non-coding sequences conserved between human and *Fugu* to be targeted as potential enhancers controlling transcription of this gene are likely to be restricted to *GLI3* introns. There, multi-species alignment of genomic sequences from 10 vertebrates (Fig. 3.1) revealed extensive conservation in closely related species, which obscured the identification of potentially functional elements embedded in intronic DNA. However, during transition from moderate (mammalian sequence comparison) to intermediate evolutionary distance (human vs. birds/amphibia) the extent of neutrally evolving sequences dropped sharply, while sequence comparison at extreme phylogenetic distance (human/teleost) reduced the number of candidates further and let us prioritize for functional analysis on 11 CNEs distributed across almost the entire *GLI3* interval encompassing conserved (minimum 60 bp at 50% identity) non-coding elements, with 2 elements each in introns 2, 3, 4, and 10 and one each in introns 1, 6, and 13 (Fig. 3.1).



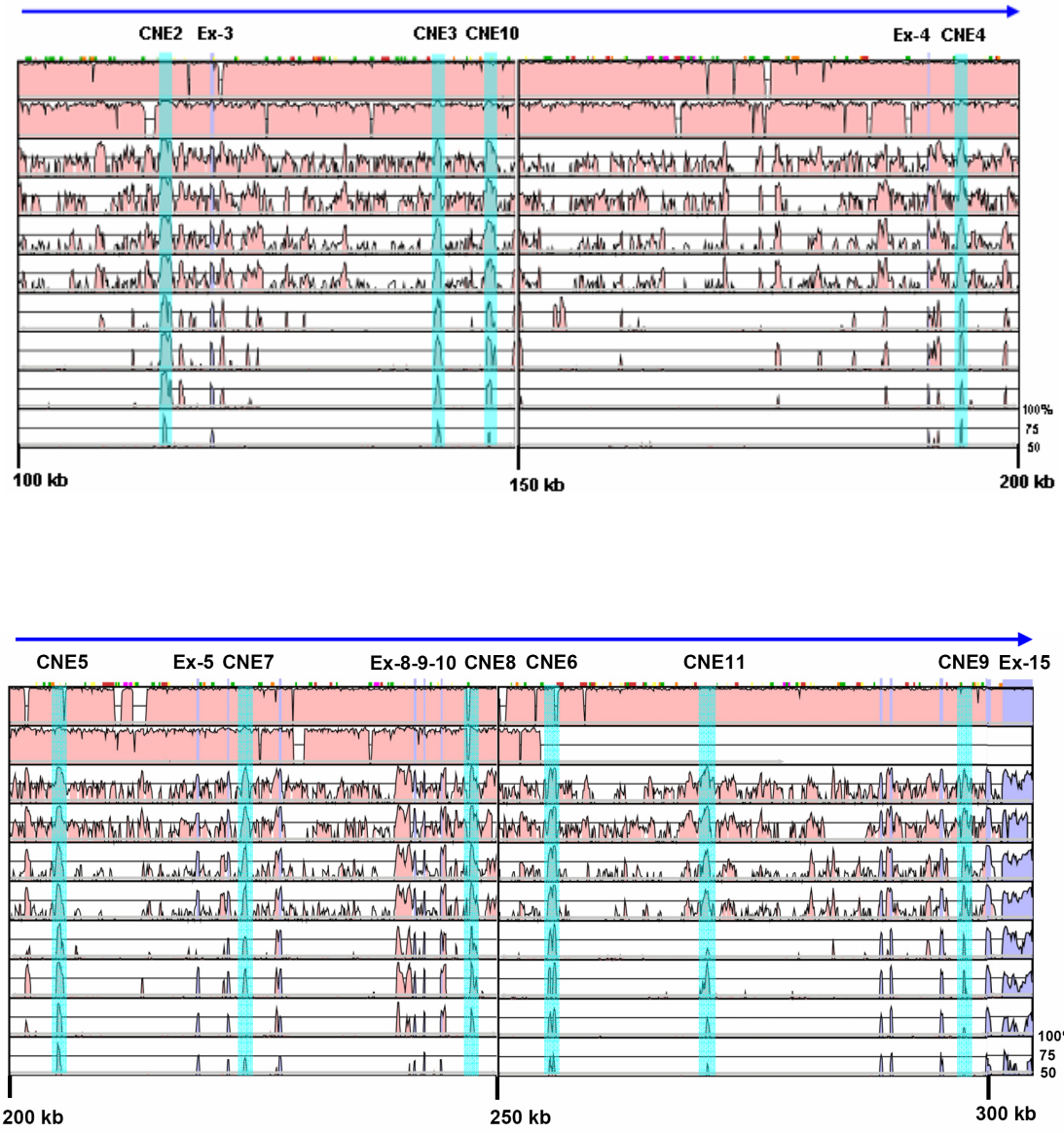


Figure. 3.1. SLAGAN alignment of the genomic region encompassing GLI3.

In each panel, human genomic *GLI3* DNA sequence from ENSEMBL is aligned with chimpanzee, rhesus, dog, cow, mouse, rat, opossum, chicken, frog, and *Fugu* orthologous regions. Alignment parameters are explained in the Methods section. CNE1-CNE11 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

The sequence comparison of human *GLI3* with other non-primate placental mammals depicts large number of conserved non-coding elements (CNEs), whereas in human vs. nonplacental-tetrapods (marsupial, monotreme, and birds) sequence comparisons the number of CNEs drops sharply (Fig. 3.2). Even though we cannot rule out the functional significance

of CNEs that appear by sequence comparison of human intronic *GLI3* sequences with orthologous sequences from placental and non-placental mammals, as evident from Fig. 3.2, we selected for functional assay the subset of intra-*GLI3* CNEs which is conserved in multiple species down to teleost fish (Table 3.1, Fig. 3.1). In addition, to test the regulatory potential of promoter proximal elements of *GLI3*, an interval of ~3.5 kb (PvuII_GLI3), encompassing a part of its minimal promoter, was also selected for in vivo functional analysis.

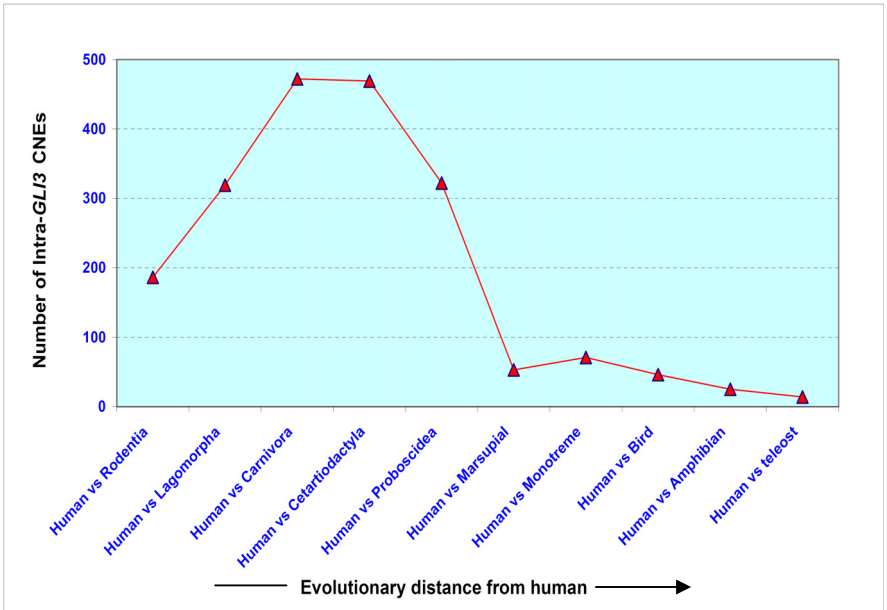


Figure. 3.2. Conservation of *GLI3* intronic intervals drops sharply in human/nonplacental tetrapods comparison.

The number of intra-*GLI3* CNEs identified with the criteria 100 bp window and 70% conservation cutoff is plotted against a pairwise comparison at various phylogenetic distances.

Table 3.1. Tetrapod-teleost Conserved Non-Coding elements (CNEs) from introns of human GLI3 selected for functional analysis

Region	Element	Amplicon coordinates Chr7	Amplicon size	Conservation human-fugu 60%; >50bp	*In vitro* activity	*In vivo* activity	Conserved putative TFBSs
Intron 1	CNE12	42239221-42239879	659 bp	190 bp	A/R	-	TBX5, PITX2, PAX6, GATA1, POU6F1
Intron 2	CNE1	42219598-42220542	945 bp	935 bp	A/R	(+)	ATF1, CDPCR1, CDXA, , EBOX, FOXM1, FOXP3, GABP, GATA1, PBX1,HOXA3, LMO2COM, MSX1, MYOGENIN, NFY, NMYC, POU3F2, USF, YY1, IRF1, AFP1, VJUN, , dHAND, SOX5, NFKB1
Intron 2	CNE2	42159050-42159483	434 bp	401 bp	-	-	CEBPDELTA, CHCH, HOX13, IRF2, LEF1B, MSX1, SP3, TCF4, EN1
Intron 3	CNE3	42131347-42131748	400 bp	378 bp	R	(–)	AREB6, ATF, EBOX , GATA1, GATA2 , GATA3 , LEF1B , LMO2COM , MYOD , NMYC , TCF4 , USF
Intron 3	CNE10	42125837-42126969	1133 bp	105 bp	A/R	(+)	CART1, CDP, CLOX, P53, E2F1, SOX5, EN1, PBX1, PITX1
Intron 4	CNE4	42079507-42079678	172 bp	160 bp	R	(–)	CREL, LEF1B, NKX25, PTF1BETA, STAT1, STAT4, STAT6
Intron 4	CNE5	42068665-42069242	578 bp	255 bp	A/R	(–)	AREB6, E2F, FREAC2, GATA1, GATA6, HNF1, HNF3 ALPHA, MEIS1, OCT1, PAX2, PBX1, PBX, TBP, XFD1
Intron 6	CNE7	42049418-42050221	804 bp	337 bp	A/R	(+)	NKX61, OCT1, POU3F2, SRY, MEF2, STAF
Intron 10	CNE6	42018164-42019025	862 bp	179 bp	A/R	(+)	OCT1, PPARA, TBX5, PBX1, PAX4, HOXD13/HOXA13, PITX1
Intron 10	CNE11	42002211-42003395	1185 bp	129 bp	A/R	(+)	SMAD3, LEF1B
Intron 13	CNE9	41975857 - 41976525	669 bp	108 bp	A/R	(+)	OCT1, PPARA, TBX5, PAX3, STAT5A

Note; Location, size, coordinates, and selected human-*Fugu* conserved transcription factor binding sites are indicated. Dual nature and exclusively repressory elements identified in transiently transfected cell cultures are represented by "A/R" (activator/repressor) and "R" symbol, respectively. The (+) sign indicates the elements which induced GFP expression in zebrafish embryos while (–) signs indicate those which could not drive GFP expression significantly. Results on CNE2 have been reported previously (Paparidis 2005).

3.2 Computational analysis to unravel within *GLI3*-CNEs highly conserved sequence patterns with potentially functional relevance

Computational investigations have revealed a wide spectrum of evolutionary constraints operating on each of the selected sub-set of CNEs. Three of them, i.e. CNE1, 2, and 5 embed highly conserved tracks. CNE6 and 10 seem to be under moderate evolutionary constraints, while CNE7, CNE9, and CNE11 depict a marginal conservation in human-fish comparison. In addition, TFBSs motif searches have been combined with phylogenetic footprinting of CNEs across different species (Loots and Ovcharenko 2004; Sandelin et al. 2004) to search within CNEs conserved putative TFBS modules (Table 3.1).

Combined application of sequence alignment and TFBSs pattern recognition computational tools, have led us to include into the study sequences with diverse degrees of conservation, not only the ultraconserved ones.

Sequence comparison among CNE1 to CNE11 did not reveal identity other than in TFBS.

The results of computational analysis with the selected subset of CNEs will be presented in the following section in order of their occurrence along the *GLI3* gene.

CNE1

CNE1 resides within intron-2 (Fig. 3.1), spanning 945bp conserved interval between human and *Fugu,* with overall identity of ~ 70%. Embedded within CNE1 there is a highly conserved sequence track of ~120 bp, almost 100% identical in human/rodents and human/bird sequence comparison and still with ~ 92% sequence identity in human/fish (Fig. 3.3).

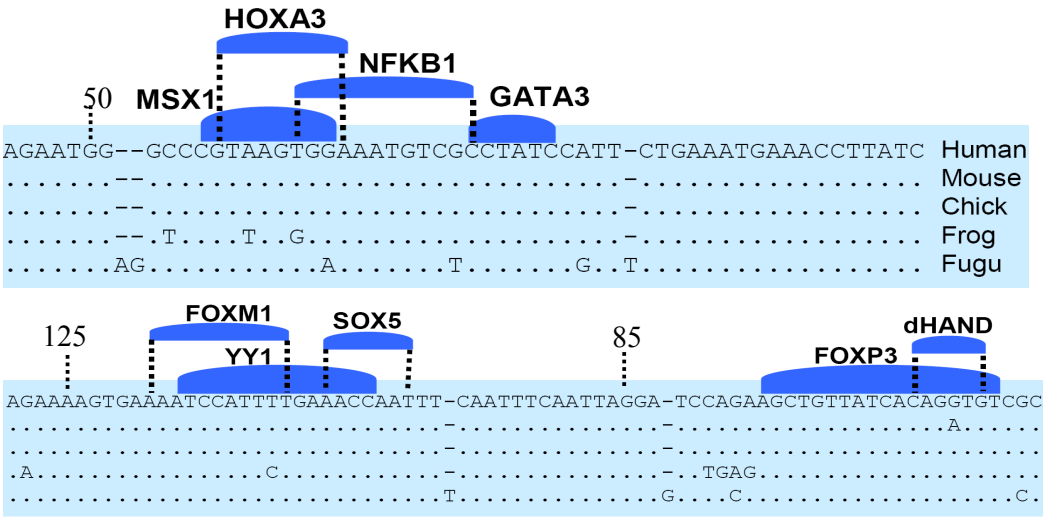


Figure. 3.3. Highly conserved sequence track in CNE1.

BLASTZ alignment of a highly conserved track within CNE1. Human/*Fugu* conserved putative binding sites for transcription factors are indicated.

In human/mouse pairwise alignment this block expands to a remarkable 230 bp (Fig. 3.4), 98% identical, with two transversion and two transition events, while in flanking sequences the conservation within CNE1 is lower within mammals and between mammals and fish lineages (65% over a 100 bp window).

```
Human   TCTCTGTGGCAGCTCTGAGTCCATTCAATATTTTCCAcaGACAGAGAAAAGTGAAAATCC
        ||||||||||||||||||||||||||||||||||||  ||||||||||||||||||||
Mouse   TCTCTGTGGCAGCTCTGAGTCCATTCAATATTTTCCAacGACAGAGAAAAGTGAAAATCC

Human   ATTTTGAAACCAATTTCAATTAGGATCCAGAAGCTGTTATCACAGgTGTCGCAGAATGGG
        |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Mouse   ATTTTGAAACCAATTTCAATTAGGATCCAGAAGCTGTTATCACAGaTGTCGCAGAATGGG

Human   CCCGTAAGTGGAAATGTCGCCTATCCATTCTGAAATGAAACCTTATCAGGgGCTACAAAG
        |||||||||||||||||||||||||||||||||||||||||||||||| |||||||||
Mouse   CCCGTAAGTGGAAATGTCGCCTATCCATTCTGAAATGAAACCTTATCAGGaGCTACAAAG

Human   ATGGCTGGGGACAGCCTATCTGTAAATTACTTTTCGTCCTGTGTTAACTCA
        ||||||||||||||||||||||||||||||||||||||||||||||||||
Mouse   ATGGCTGGGGACAGCCTATCTGTAAATTACTTTTCGTCCTGTGTTAACTCA
```

Figure. 3.4. Pairwise alignment shows a human/mouse ultraconserved sequence track within CNE1.

The human/fish highly conserved track within CNE1 (of 92% identity, indicated by red letters) shows in human/mouse comparison an expansion of the extreme level of evolutionary constraints along its both sides.

Human/*Fugu* pairwise sequence comparison was used to search for TFBSs that were conserved over such an extreme phylogenetic separation, i.e. 450 Mya (Table 3.1). Computer programs ConSite (Sandelin et al. 2004) and rVISTA v 2.0 (Loots and Ovcharenko 2004) have been used to explore the JASPAR and TRANSFAC library simultaneously under the criteria given in the Methods section.

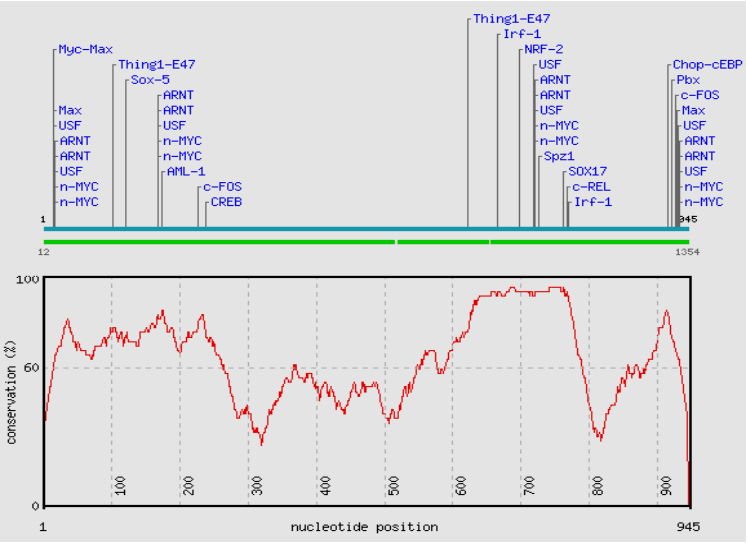


Figure. 3.5. Consite comparison of human CNE1 sequence with *Fugu* to identify conserved TFBSs.

The plot was generated with 70% conservation cutoff, 50bp window size, and 75% transcription factor binding site score threshold settings. Identified TFBS are shown above. The y-axis in the graph specifies the percentage of identical nucleotides. The x-axis refers to the nucleotide position in the human sequence.

From the predicted conserved TFBS identified with Vista v 2.0 and Consite (Fig. 3.5), the ones shown in Table 3.1 are those whose binding factors are co-expressed with *GLI3*. For instance, CNE1 contains a tetrapod/teleost conserved putative binding site (CGTAAGTC) for the developmentally important transcription factor MSX1 know to be involved in the development of teeth (Vastardis et al. 1996), CNS, heart (Campbell et al. 1989), and limb (Hu et al. 1998). A binding site for HOXA3 overlaps with MSX1. That developmentally important transcription factor provides cells with specific positional identities on the anterior/posterior axis, and it is expressed in heart, limb and spinal cord (Conway et al. 1997). Intra-CNE1 phylogenetic footprinting also revealed a conserved putative binding site for the SRY- related HMG-box gene SOX5, which is expressed in brain, testis, kidney, lungs, and is also involved in skeletal development (Smits et al. 2001). A conserved binding site for NFKB1 (GGAAATGTCG) has also been identified, which is known to play a role in skeletal, (Iotsova et al. 1997), lung, and heart development (Kanters et al. 2003), and is also found in kidney, liver, and pancreas (http://www.genecards.org). The clusters of human/fish conserved TFBSs for developmentally important transcriptional regulators within CNE1 (Table. 1.1 and Fig. 3.5) depict a putative ancient regulatory architecture which remained essentially un-altered, even at the individual TFBS level, over the course of 450 Mya.

CNE10

CNE10 showed a moderately conserved tetrapod/teleost track of 106 bp (Fig. 3.6) within intron 3 of *GLI3*.

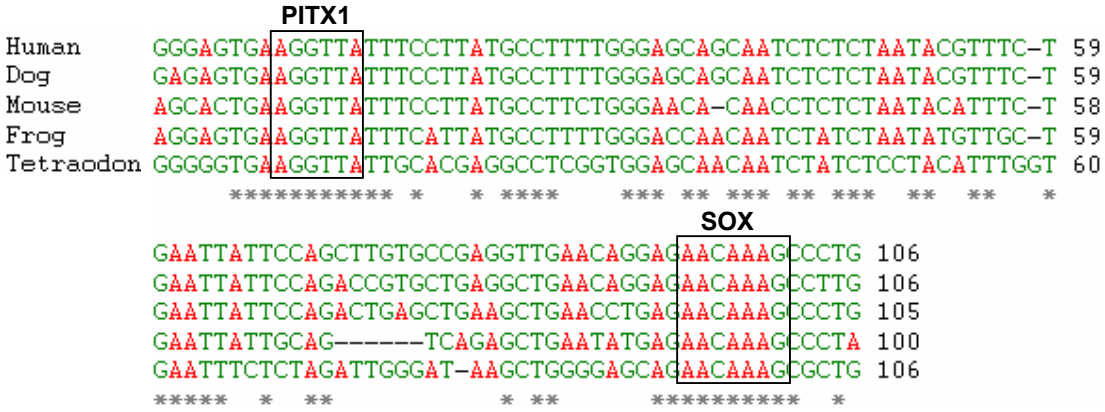


Figure. 3.6. CNE10 shows moderate conservation in multispecies alignment.

Multiple alignment of human, mouse, dog, frog, tetraodon CNE10 sequences with CLUSTALW. Space in between the alignment show deletion events, while a star symbol underneath represents a nucleotide position which is conserved in all lineages.

Search for human/*Fugu* conserved TFBSs within CNE10 revealed putative binding sites for several developmentally important transcription factors (Table 3.1).

CNE5

Human/*Fugu* pairwise alignment of the CNE5 exhibited 85% identity over 255bp. The human/fish conserved fragment shows clustering of conserved TFBSs (Fig. 3.7)

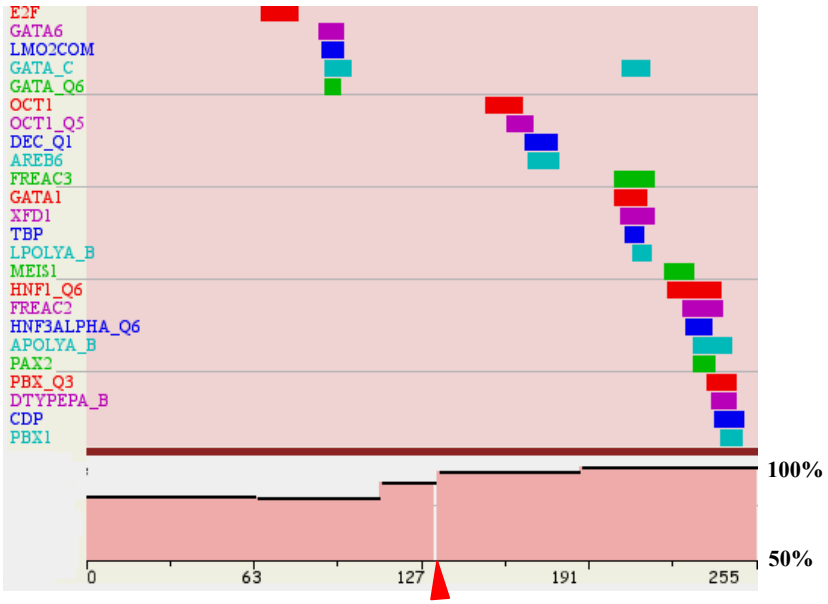


Figure. 3.7. Conserved transcription factor-binding sites in CNE5.

The 255 bp human-*Fugu* conserved element was analyzed for the presence of conserved TFBSs using rVISTA 2.0, with a matrix similarity cutoff at 0.85. Alignment between *Fugu* and human sequences is depicted in blocks ranging from 50 to 100% conservation. Positions of conserved putative TFBSs are indicated by colored boxes above the alignment. The red arrowhead underneath the alignment plot indicates a deleted region in the human sequence.

Multiple sequence alignment indicated a deletion event (Fig. 3.8) of 37 bp shared by all tetrapods, while the corresponding fragment was still conserved in teleosts, i.e. the *Fugu*/tetraodon comparison. The deletion event might happened somewhere in between the transition from fish to amphibian lineage. Tolerance for such a big deletion within the core of a putative enhancer element in the tetrapod lineage (from mammals to amphibians), while the remaining CNE5 interval is highly constraint in all tetrapods (and probably functional) (Fig. 3.8), suggests that loss of this fragment from CNE5 might have led to a pivotal functional difference between the teleost and tetrapod lineages.

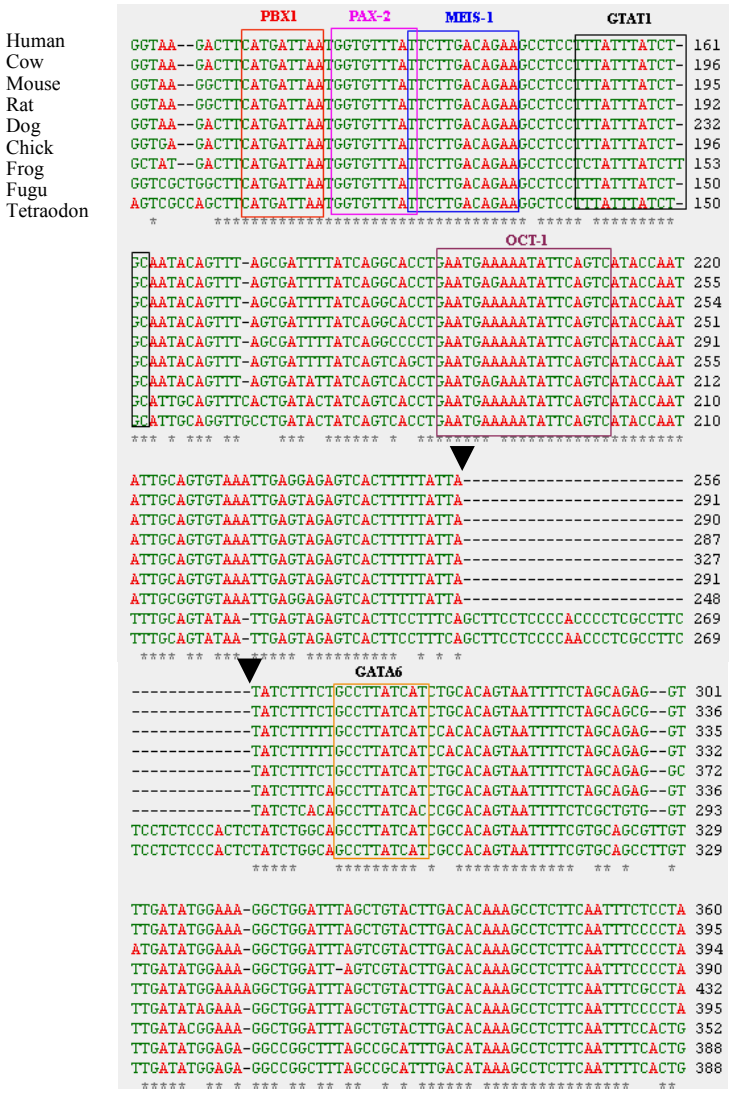


Figure. 3.8. Tetrapod specific deletion of a 37 bp fragment in CNE5

Multiple alignment of human, cow, mouse, rat, dog, chick, frog, *Fugu,* and tetraodon CNE5 sequences with CLUSTALW. Spaces in between the alignment show the deletion events, while a star symbol underneath represents nucleotide positions which are conserved in all lineages. Conserved TFBS modules are enclosed within boxes. Arrowheads indicate the 37bp deletion event shared by tetrapodes. A subset of the binding sites including the deletion event depicted graphically in Fig. 3.7 is shown at nucleotide level in this Figure.

Furthermore, binding sites for developmentally important homeobox (MEIS1 and PBX1) and paired box (PAX2) members of transcription factors are remarkably conserved (Fig. 3.8) and exist in form of unaltered contiguous blocks. PBX1, MEIS1, and PAX2 are co-expressed with GLI3 in limbs (Mercader et al. 1999), CNS, spinal cord, liver, kidney, eye (Tellier et al. 2000), and pancreas (Kim et al. 2002), which validates the notion of these putative conserved TFBS in CNE5 being involved in endogenous *GLI3* expression.

CNE6

Residing within intron-10 of *GLI3,* CNE6 encompasses a 179 bp region of overall sequence conservation of 73.7% in human/*Fugu* comparison (Fig. 3.9) and features conserved binding sites for developmentally important transcription factors (Table 3.1).

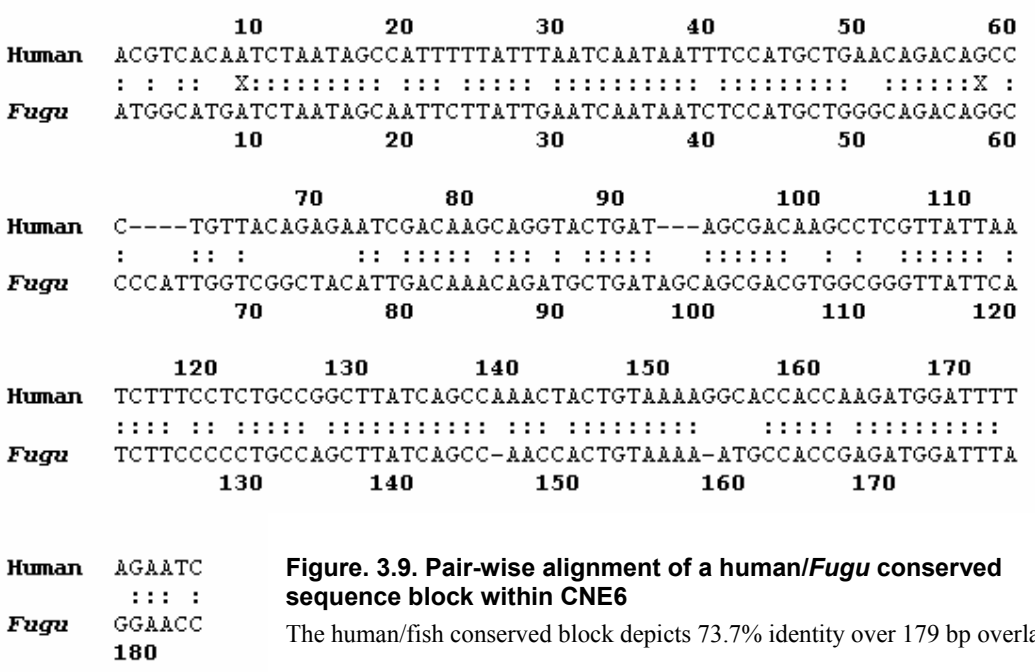
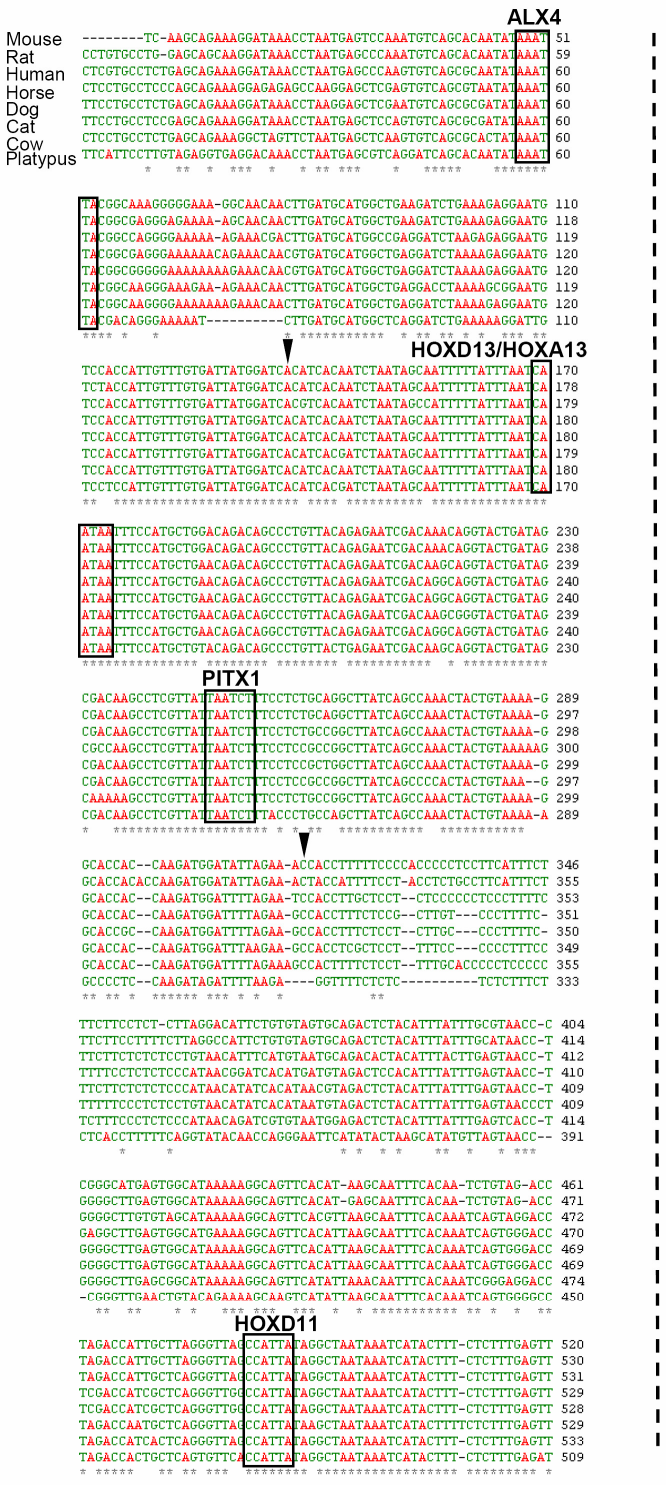


Figure. 3.9. Pair-wise alignment of a human/*Fugu* conserved sequence block within CNE6

The human/fish conserved block depicts 73.7% identity over 179 bp overlap.

While the conservation of human CNE6 down to the bony fish lineage is confined to a limited sequence block, the element extends to 862bp in human versus placental/non-placental mammalian sequence comparison (Fig. 3.10). Phylogenetic footprinting within the sequence area of CNE6 conserved in mammals revealed highly conserved binding motifs (down to monotremes) for numbers of developmental regulators that are known to be co-expressed with *GLI3* (Fig. 3.10).



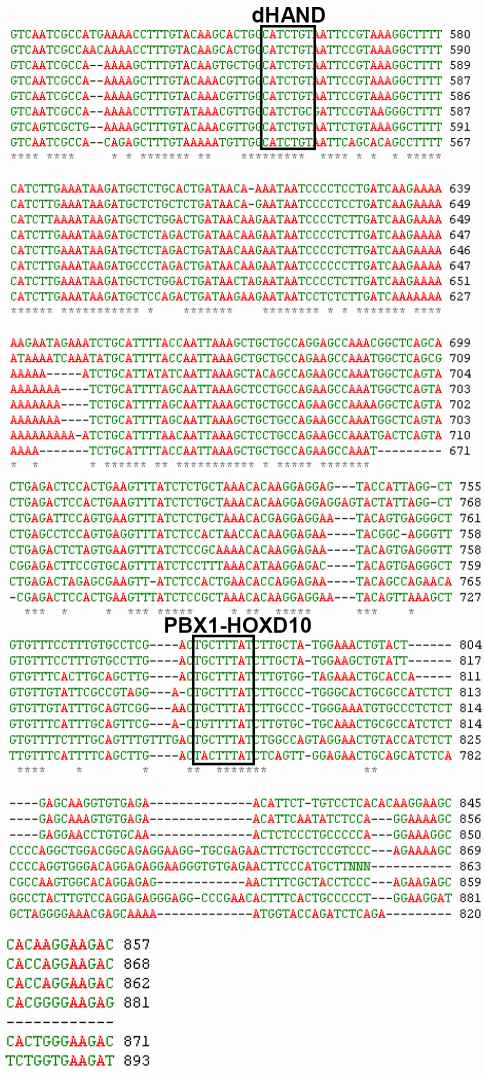


Figure. 3.10. Conservation of the human CNE6 interval down to non-placental mammals (monotremes)

The CNE6 interval expands along both edges of a human/fish conserved block (bracketed by black arrowheads) up-to remarkable 862 bp in multi-species sequence comparison down to non-placental mammalian lineage. The phylogenetic footprinting revealed conserved binding motifs for numbers of developmental regulators that are known to be co-expressed with *GLI3* during early embryonic development.

CNE11

The 1185 bp CNE11 sequence is positioned in the neighborhood of CNE6 (~ 16kb apart), within intron-10 of *GLI3*. It harbors a sequence core conserved down to the fish lineage (black arrowheads in Fig. 3.11). Sequence alignment involving multiple mammalian species revealed areas of high conservation encompassing binding sites for early developmental regulators that are known to be co-expressed with GLI3 during early mammalian embryonic development, e.g. HOXD13, dHAND and TBX3 (Fig. 3.11).

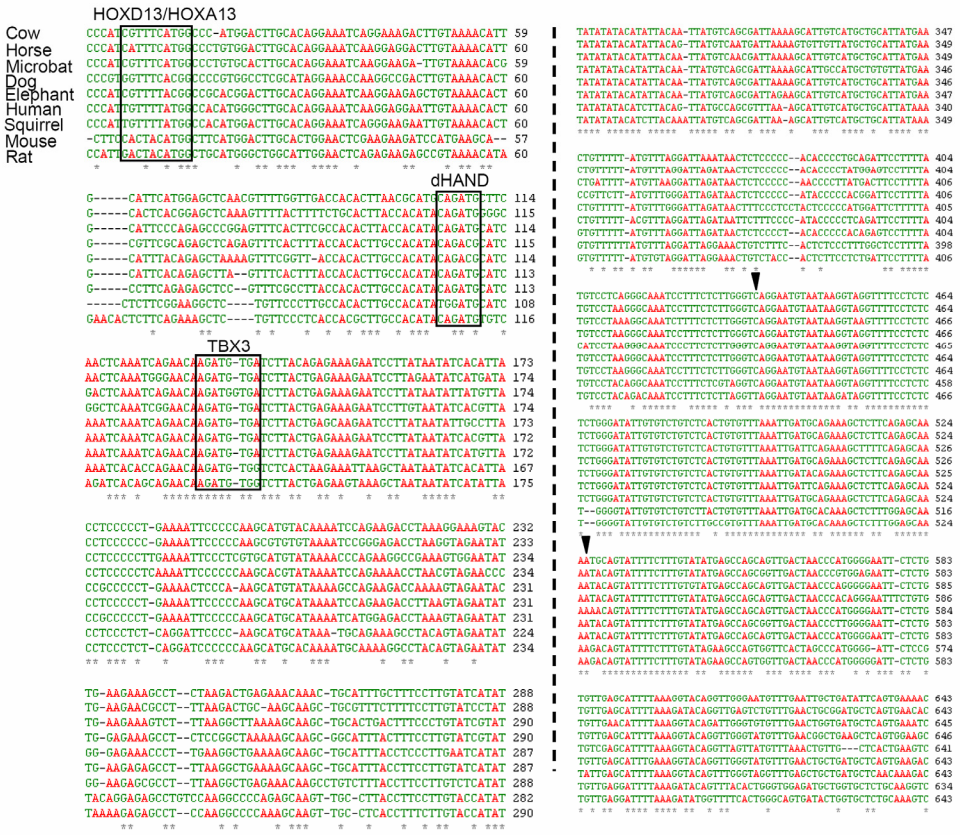


Figure. 3.11. Part of human CNE11 sequence (643bp/1185bp) indicting high conservation in multiple mammalian species

Binding sites of early developmental regulators that are known to be co-expressed with GLI3 during early mammalian development are shown. The black arrowheads indicate a region within CNE11 that is marginally conserved down to the fish lineage.

CNE9

CNE9, positioned within intron-13 of *GLI3,* exhibits a small size (~30bp) fragment conserved down to the teleost lineage (Fig. 3.12). Although this element is relatively small, it is remarkably highly conserved, suggesting a critical functional role.

```
Human      CTTCCTAAGGACAGTGGGTAATAAGGTGAG
Dog        CTTCCTAAGGACAGTGGGTAATAAGGTGAG
Mouse      CTTCCTAAGGACAGTGGGTAATAAGGTGAG
Rat        CTTCCTAAGGACAGTGGGTAATAAGGTGAG
Chick      CTTCCTAAGGACAGTGGGTAATAAGGTGAG
Opossum    CTTCCTAAGGACAGTGGGTAATCAGGTGAG
Fugu       CTTGCTAAGGACAGTGGGTAATCAGGTGAG
Tetraodon CTTACTAAGGACAGTGGGTAATCAGGTGAG
           *** ****************** *******
```

Figure. 3.12. Over a short interval CNE9 depicts exceptionally high conservation down to fish.

Multiple alignment of human, dog, mouse, rat, chick, opossum, *Fugu,* tetraodon CNE9 sequences with CLUSTALW. A star symbol underneath represents nucleotide positions conserved in all lineages.

3.3 *In-vitro* functional analysis of intra-*GLI3* conserved non-coding elements

In order to test the selected subset of sequence elements for their potential to regulate reporter gene expression, recombinant constructs with CNEs placed in either orientation upstream of a luciferase gene controlled by the heterologous SV40 promoter or the human minimal *GLI3* promoter (Fig. 3.13A) were transiently transfected into two human cell lines. The H661 cell line expresses endogenous *GLI3* whereas H441 does not express this gene. In dual luciferase assays eight elements (CNE 1, 5, 6, 7, 9, 10, 11, and 12) showed in H661 cells activating potential whereas two elements (CNE3 and CNE4) repressed reporter gene expression below the level achieved by either promoter alone (Fig. 3.13B).

In the H441 cell line in which GLI3 is not expressed, not only the previously repressing elements CNE3 and CNE4 but also CNE 1, 5, 6, 7, 9, 10, 11, and 12 which had shown activating potential in endogenous GLI3 positive context (H661) exhibited a strong repressing activity (Fig. 3.13C). Thus, the cell based reporter assay identified two categories of intra-*GLI3* regulatory elements, context independent repressors and enhancers with a context dependent dual nature, serving as activators in GLI3 positive context and as repressors in cells without endogenous GLI3 expression.

The functional data from the H661 and H441 cell lines (Fig. 3.13) advocates the *GLI3* specific regulatory activity of intra-*GLI3* conserved elements.

3.3.1 Transcriptional silencers

CNE3

Residing within intron 3, the 400 bp element (Table 3.1) represses the luciferase expression from the minimal *GLI3* promoter by ~ 40% and 37% in GLI3 positive (Fig. 3.13B) and negative context

(Fig. 3.13C), respectively, when compared to a control construct. The same element suppressed the transcription from the heterologous SV40 promoter by ~90% in both H661 and H441 cell lines.

CNE4

CNE4 residing within intron 4 (Fig. 3.1) was the second intra-*GLI3* tetrapod-teleost conserved non-coding element which showed a repressing potential towards both minimal *GLI3* and SV40 promoters in a context independent manner (Fig. 3.13B-C).

3.3.2 Context dependent dual nature (activator / silencer) elements

CNE12

In GLI3 positive context CNE12 up-regulates transcription by 3-fold only in combination with an SV40 promoter (Fig. 3.13B). In GLI3 negative context CNE12 indicated a repressing trend towards both homologous and heterologous promoters (Fig. 3.13C).

CNE1

CNE1 induced the reporter gene activity in combination with both the heterologous SV40 and the homologous minimal *GLI3* promoter. In the H661 cell line it increased the SV40 activity by 10-fold and the minimal *GLI3* promoter activity by ~9-fold when compared to the control vectors (Fig. 3.13B).

With the H441 cell line, CNE1 shows repressing potential (Fig. 3.13C) by decreasing the heterologous and homologous promoter activity by 60% and 50%, respectively, when compared to control constructs.

CNE10

When tested in the H661 cell line (Fig. 3.13B), CNE10 up-regulates the luciferase gene expression by 4-fold and 6-fold from the minimal *GLI3* and SV40 promoter, respectively, when compared to control vectors.

CNE10 containing constructs revealed a repressing trend towards the minimal *GLI3* and the SV40 promoter in GLI3 negative context (Fig. 3.13C) by decreasing the reporter gene expression by 64% from GLI3 minimal promoter and 32% from SV40 promoter.

CNE5

Positioned within intron-4 of *GLI3,* this human/fish conserved element in H661 cells (Fig. 3.13B) drove luciferase expression by 2.3-fold and 1.9-fold in combination with SV40 and minimal *GLI3* promoters, respectively, when compared with the control constructs.

This element, when tested in H441 cell line (Fig. 3.13C), indicated a repressing potential and reduced the SV40 and minimal *GLI3* promoter activity by 62% and 75%, respectively.

CNE7

An amplicon of 804 bp from intron-6 of *GLI3* induced luciferase gene expression by 5-fold only in combination with the SV40 promoter in a direction dependent manner, while in the context of a *GLI3* promoter the CNE7 containing construct did not induce the reporter gene expression to above the level of the control construct (Fig. 3.13B).

CNE6

When transiently transfected into the H661 cell line (Fig. 3.13B), CNE6 containing constructs demonstrated an increased reporter gene expression by 10-fold and 7-fold in context of the SV40 and the minimal GLI3 promoter, respectively, when compared with the control constructs where the luciferase activity was driven alone by either the SV40 or the minimal *GLI3* promoter.

In GLI3 negative context, i.e. the H441 cell line, CNE6 showed a repressing potential towards both minimal GLI3 and SV40 promoters by decreasing the reporter gene expression ~50%.

CNE11

In GLI3 positive context, CNE11 up-regulates transcription by 16-fold only when luciferase expression was driven by the SV40 promoter.

In a GLI3 negative context, CNE11 depicted a repressing trend, towards both homologous and heterologous promoters.

CNE9

In context of the SV40 promoter, CNE9 enhanced the reporter gene expression by 12.4-fold (Fig. 3.13B), while it did not show any activating potential in combination with the *GLI3* promoter when assayed within a GLI3 expressing endogenous environment.

In the GLI3 negative context (Fig. 3.13C), CNE9 repressed the basic transcription level achieved by either of the promoters alone.

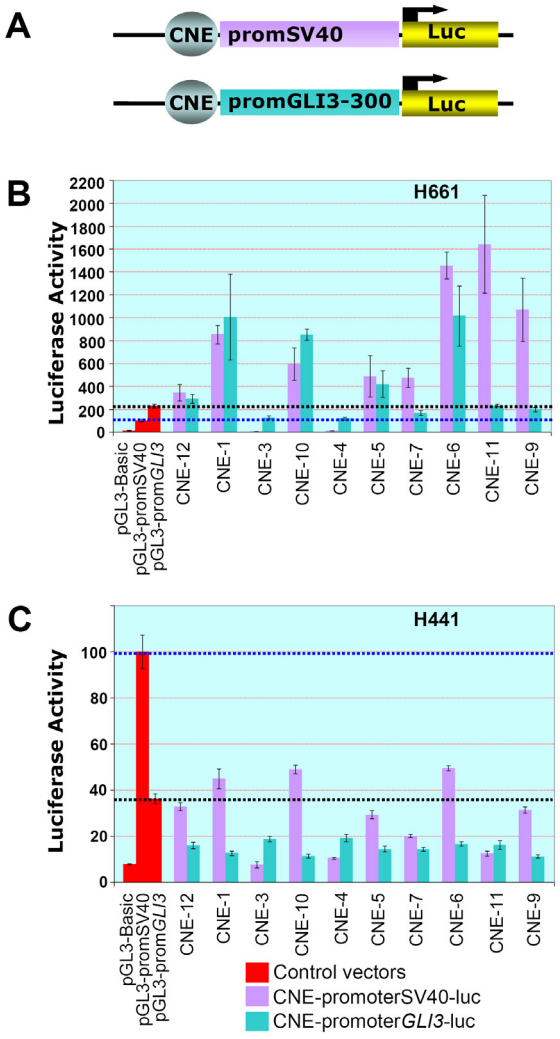


Figure. 3.13. CNEs regulate luciferase reporter gene expression in transiently transfected human cell lines.

(A) Diagrams of the reporter constructs employed to test the regulatory potential of highly conserved non-coding sequence elements (CNEs) from introns of human *GLI3*. CNEs were associated in the pGL3-Basic vector with a minimal *GLI3* promoter (pGL3-CNE-prom*GLI3*-300-luc) or a heterologous SV40 promoter (pGL3-CNE promSV40-luc) upstream of a firefly luciferase gene.
(B) Luciferase activity of reporter constructs in human H661 cells expressing endogenous *GLI3*.
(C) Luciferase activity of reporter constructs in human H441 cells not expressing endogenous *GLI3*.
The pGL3-Basic vector, which lacks the promoter/enhancer inserts was used as a negative control. Luciferase activity in cells transiently transfected with the positive control, a construct containing a SV40 promoter upstream of the reporter gene (pGL3-promSV40-luc), was taken as 100% (blue dotted line). A plasmid expressing *Renilla* luciferase was co-transfected as a standard for transcription efficiency. Average firefly luciferase reporter activities relative to *Renilla* luciferase activity from 3 triplicate transfection experiments are depicted as percentage of activity obtained with the positive control vector (B, C). The standard errors of the mean are shown. Black dotted lines indicate the luciferase expression level reached in each cell line with the pGL3-prom*GLI3*-300-luc vector

3.4 *In-vitro* deletion analysis of a selected sub-set of CNEs

In order to define functionally critical regions within CNEs and to understand the significance of strength of evolutionary constraints on defining their overall activity, we prioritized three elements CNE1, CNE5, and CNE6 for dissection and subsequent analysis of the fragments by transient transfection assays in H661 cells.

CNE1

CNE1 spans a 945 bp human/fish conserved track with overall human/fish sequence similarity of ~71%. Close inspection of CNE1 revealed a core sequence block of ~125 bp (*hcCNE1-125 bp*) under severe evolutionary constraints almost unaltered in human/mouse and human/chick sequence comparison, while depicting ~92% sequence identity in human/*Fugu* comparison (Fig. 3.14A).

In order to test the functional significance of the *hcCNE1-125 bp* track, we generated two different deletion constructs. One contained *hcCNE1-125 bp* alone upstream of the minimal *GLI3* promoter without the flanking less conserved region. In the second construct the *hcCNE1-125 bp* fragment had been deleted from wild type CNE1 to investigate the activity of the flanking human/*Fugu* moderately conserved region without the core element (Fig. 3.14B).

Deletion of the 125 bp highly conserved track reduced the activity by 67% compared to the wild type construct, but the deleted CNE1 without *hcCNE1-125 bp* was still able to induce reporter gene expression 4-fold compared to control vector where luciferase expression was driven alone by the minimal *GLI3* promoter. Tested alone, *hcCNE1-125 bp* was unable to show any activating potential (Fig. 3.14C).

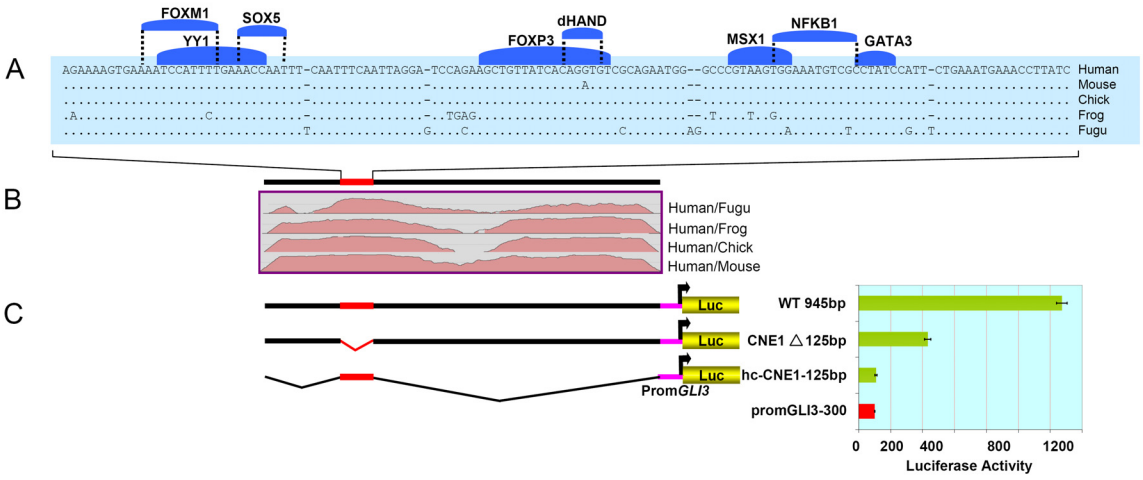


Figure. 3.14. Deletion analysis reveals a critical role of *hc-CNE1-125 bp* for the regulatory potential of CNE1.

(A) BLASTZ alignment of highly-conserved 125 bp sequence fragment embedded within CNE1 from human, mouse, chick, frog and *Fugu* is presented along with predicted conserved TFBSs.
(B) CNE1 alignment plot for human, mouse, chick, frog and *Fugu* sequences by using human sequence as base line.
(C) Architecture of CNE1 wild type and deletion constructs; red portion of bar depicts the highly conserved part, and the black portions indicate the flanking less conserved segments. Luciferase activity obtained in H661 cells after transient transfection of reporter constructs is shown in the diagram at the right side. Reporter gene expression is driven by CNE1 fragments upstream of the human *GLI3* minimal promoter. The red bar depicts luciferase expression (100%) in H661 cells driven alone by the control *GLI3* minimal promoter (Prom-*GLI3*-300), while green bars represent the activity recorded for the vectors containing experimental reporter constructs, i.e. wild type CNE1 (WT 945 bp), CNE1 with deleted highly conserved track (CNE1Δ125 bp), and highly conserved 125 bp fragment alone (*hc*CNE1-125 bp). The standard errors of the mean are shown.

CNE5

CNE5 harbors two highly conserved blocks interrupted by a short less conserved fragment (Fig. 3.15B). Phylogenetic footprinting reveals contiguous binding sites for three developmentally important homeobox and paired box transcription factors within one of these blocks , PBX1, PAX2 and MEIS1, 100% conserved in multispecies sequence comparison from human to fish. The PAX2 and MEIS1 binding sites overlap by one nucleotide.

A 50 bp module encompassing the PBX1, PAX2, and MEIS1 binding sites was excised from wild type CNE5 fragment. The resulting deleted CNE5 and a 144 bp 90% human/*Fugu* conserved track encompassing both highly conserved blocks were cloned into the pGL3-promGLI3-300-Luc and compared to wt CNE5 for their potential to enhance reporter gene expression when transiently transfected into H661 cells. In contrast to wt CNE5, both elements were unable to activate basic expression (Fig. 3.15B and 3.15C).

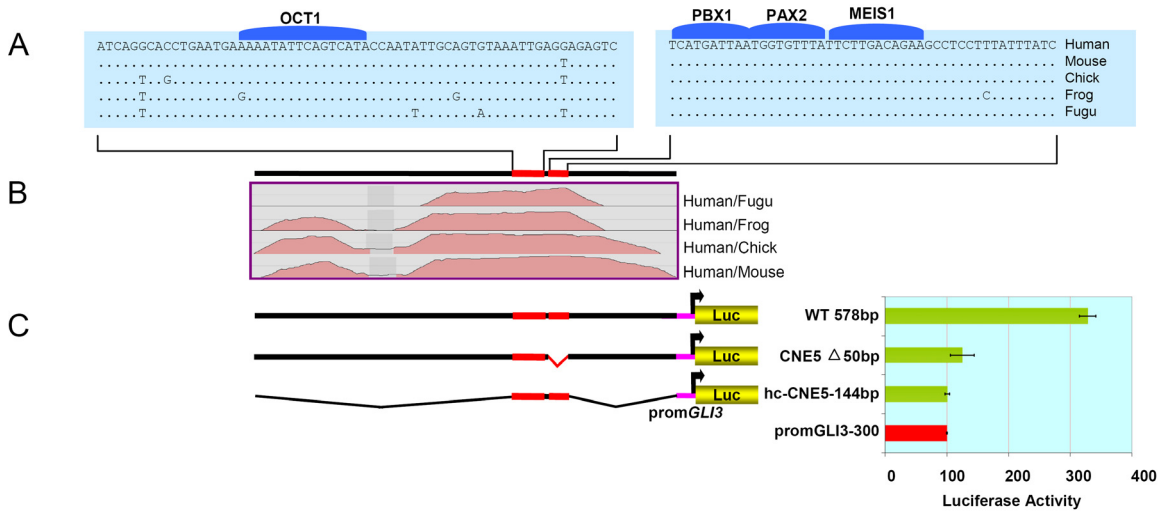


Figure. 3.15. Putative binding sites for individual trans-acting factors are necessary but not sufficient for the activating potential of CNE5.

(A) BLASTZ alignment of highly conserved fragments embedded within CNE5 along with predicted conserved TFBSs.
(B) CNE5 alignment plot of human, mouse, chick, frog and *Fugu* sequences by using human sequence as base line.
(C) Architecture of wild type and deletion constructs; red portion of bar depicts the highly conserved human/fish segments. Luciferase activity obtained in H661 cells after transient transfection of reporter constructs is shown in the diagram at the right side. Reporter gene expression is driven by CNE5 fragments upstream of the human *GLI3* minimal promoter. The red bar depicts luciferase expression (100%) in H661 cells driven alone by the control *GLI3* minimal promoter (Prom*GLI3*-300), while green bars represent the activity recorded for the vectors containing experimental reporter constructs, i.e. wild type CNE5 (WT 578 bp), CNE5 with deleted PBX1, PAX2 and MEIS1 binding module (CNE5Δ50 bp), and 144bp fragment (hcCNE5-144 bp). The standard errors of the mean are shown. Deletion of the 50 bp fragment almost entirely extinguishes the strong activating potential of CNE5. The isolated 144 bp fragment can not activate expression.

CNE6

The wild type 862 bp CNE6 fragment shows overall 87% sequence identity in human/mouse comparison and contains a human/*Fugu* moderately conserved track of 179 bp length. This track encompasses a 35 bp highly conserved site. In order to investigate the participation of the 179 bp human/fish and the flanking tetrapod conserved elements to the over-all *in vitro* enhancer activity of wt CNE6, each region was investigated separately (Fig. 3.16C).

The deletion of the 179 bp element reduced the activity of CNE6 by ~70% compared to wild type construct. However, compared to the control vector where transcription was driven only by the minimal *GLI3* promoter, this deleted CNE6 was still able to up-regulate the reporter gene expression by 1.7-fold.

Insertion of the 179 bp fragment upstream of the minimal *GLI3* promoter did not result in up-regulation of reporter gene expression compared to control (Fig. 3.16B and 3.16C).

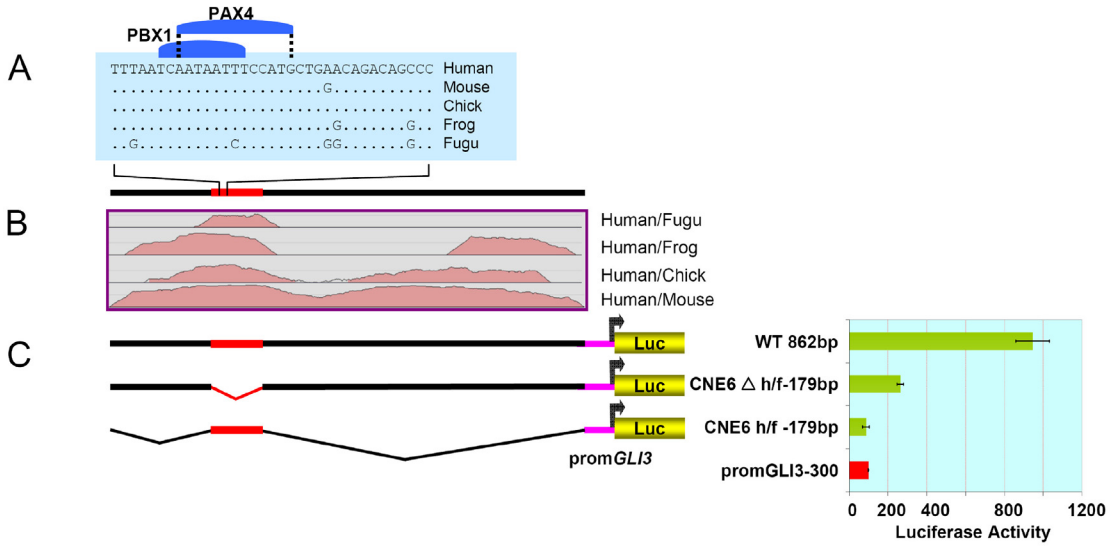


Figure. 3.16. CNE6 sequences flanking human/fish conserved track show residual enhancer activity.

(A) BLASTZ alignment of the highest conserved 40 bp along with two predicted conserved TFBSs from the human/*Fugu* conserved block within CNE6.
(B) CNE6 alignment plot of human, mouse, chick, frog and *Fugu* sequences by using human sequence as base line.
(C) Architecture of wild type and deletion constructs; red portion of bar depicts the highly conserved human/fish segment. Luciferase activity obtained in H661 cells after transient transfection of reporter constructs is shown in the diagram at the right side. Reporter gene expression is driven by CNE5 fragments upstream of the human *GLI3* minimal promoter. The red bar depicts luciferase expression (100%) in H661 cells driven alone by the control *GLI3* minimal promoter (Prom*GLI3*-300), while green bars represent the activity recorded for the vectors containing experimental reporter constructs, i.e. wild type CNE6 (WT 862 bp), CNE6 with deleted human/*Fugu* conserved block (CNE6Δh/f-179bp), and the 72% human/fish conserved fragment (CNE6h/f-179 bp). The standard errors of the mean are shown. CNE6Δh/f-179 bp can still enhance reporter gene transcription more than two-fold. The isolated 179 bp fragment can not activate expression.

3.5 *In-vivo* functional analysis of CNEs with transiently transfected zebrafish embryos

The *in vitro* strategy was successful enough to resolve the dual nature, i.e. context dependent activator/repressor potential of eight intra-*GLI3* conserved non-coding elements, and it identified also two elements with only repressing behavior towards both the minimal *GLI3* and SV40 promoters. In order to address the *in vivo* role of this subset of *GLI3* associated tetrapod-teleost conserved non-coding elements (CNEs 1, 2, 3, 4, 5, 6, 7, 9, 10, 11,) we selected a medium throughput strategy (Woolfe et al. 2005), i.e. transient reporter gene expression from a human β-globin promoter under the

influence of each putative enhancer element in zebrafish embryos. This approach has recently shown its immense potential to assign with reproducibility and speed a functional role to conserved non-coding elements by exploiting the transparency of developing zebrafish embryos (Goode et al. 2005; McEwen et al. 2006; Muller et al. 1999; Woolfe et al. 2005). In addition to tetrapod-teleost conserved elements (CNEs), a ~ 3.5 kb sequence element upstream of exon-1 (PvuII_GLI3) including a part of promoter has also been subjected to *in vivo* assay. Zebrafish enhancer / GFP reporter assay was performed in collaborations with the lab of Professor Greg Elgar, QMUL, UK,

Because of the high mosaic nature of *in vivo* transient reporter expression, a large number of embryos has been microinjected and screened at a time (mean of 209 embryos for each element), to derive an over all impression of enhancer activity at different locations for each element through cumulative data from all GFP expressing embryos.

Consistent with the results of the *in vitro* assay, all but one activator elements (CNE5) induced GFP expression in a significant proportion of microinjected zebrafish embryos, while none of the *in vitro* identified repressors (CNE3 and CNE4) were able to show detectable GFP induction. In addition, one of the elements, i.e. CNE2, to which we could not assign any functional role either as an activator or repressor with the cell based reporter assay, consistently induced *in vivo* GFP-expression in a significant proportion of microinjected zebrafish embryos.

GLI3_PvuII

In order to test the regulatory potential of an immediate up-stream interval of *GLI3* (promoter proximal regions), an *in vivo* assay has been performed with a fragment extending from nucleotide six of the untranslated exon 1 to basepair +3503 bp upstream of *GLI3*. This ~3.5 kb fragment overlaps for 198 bp with the minimal *GLI3* promoter (425 bp) (Paparidis 2005) and is not conserved in tetrapod/teleost sequence comparison, however it did show conservation within tetrapod orthologous intervals (Fig. 3.17).

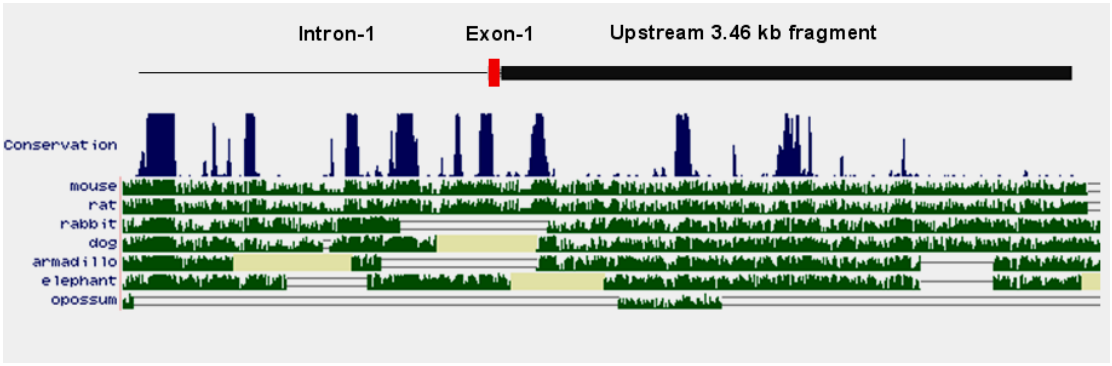


Figure. 3.17. The immediate upstream interval of *GLI3* is conserved only in tetrapods.

BLASTZ alignment of mouse, rat, rabbit, dog, armadillo, elephant, and opossum orthologous intervals by using human as a base line sequence. Shown here is a part of *GLI3*-intron-1, exon-1 and the immediate upstream ~3.5 kb GLI3_PvuII fragment. Green plots depict a pairwise comparison, while the blue plot above represents the overall conservation view. Yellow bars within the green plots indicates incomplete nucleotide information.

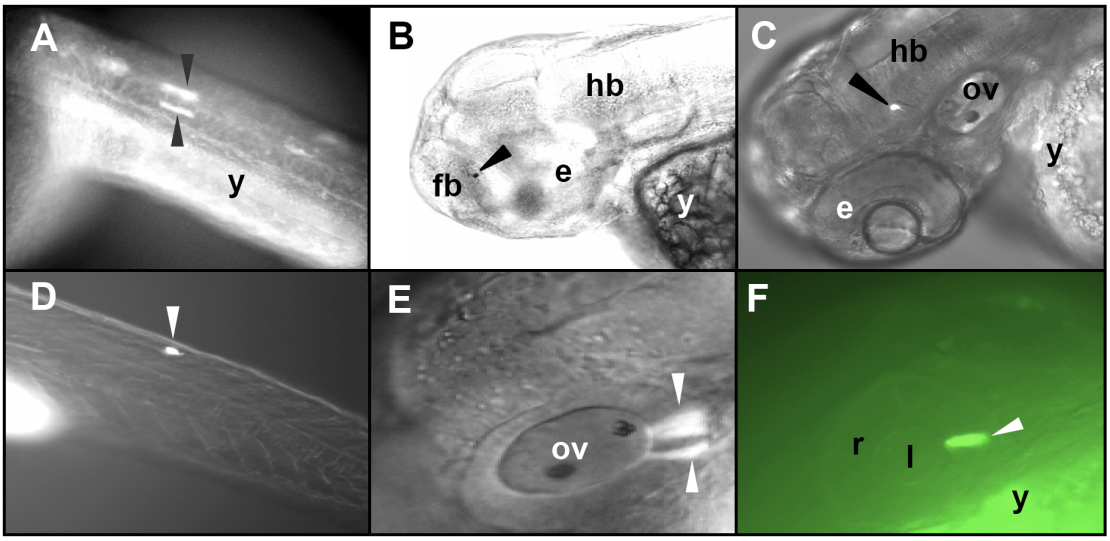


Figure. 3.18. The *GLI3* promoter proximal region shows widespread activity in zebrafish

GFP expression is shown in fixed tissues after wholemount anti-GFP immunostaining, bright field view (B), or GFP fluorescence is shown in live embryos, combining bright field and fluorescence microscopy (A, C, D, E, F). Black and white arrowheads indicate GFP expressing cells. Presented here are GFP expressing (A) muscle cells in the trunk region, (B-D) neurons in diencephalon, hindbrain (rhombencephalon) near the caudal end of spinal cord (E) GFP is expressed in otic epithelium just below the posterior otolith (F) retinal ganglion cell. e, eye; fb, forebrain; hb, hindbrain; l, lens; mb, midbrain; ov, otic vesicle; r, retina; y, yolk.

The GLI3_PvuII mediated GFP expression was not restricted to one particular region. Activity was observed in CNS (36% of expressing embryos 3.18 B-D) most prominently in the hindbrain and spinal cord (14% of expressing embryos). Additionally, the reporter activity was detected in blood cells (17% of expressing embryos), the pericardial region (35% of expressing embryos), sensory organs (10% of expressing embryos, Fig. 3.18E and F), skeletal muscle (11% of expressing embryos, Fig. 3.18A), and also in the skin and median fin fold (23% and 17% of expressing embryos, respectively).

CNE1

CNE1 directed GFP expression (Fig. 3.19) prominently in various subdivisions of CNS (Fig.3.19D-F), i.e. forebrain, midbrain and hindbrain with 22%, 32% and 58% of expressing embryos, respectively, in day two of development (~26-33 hpf). Most notable, the GFP expression with CNE1 was also observed within the cardiac chambers, where the GFP expression was most frequent on day 3 of development (~50-54 hpf) (30% expressing embryos, Fig. 3.19Aa and Ab) as compared to day two (10% of expressing embryos). In addition to CNS and heart, GFP expression was also observed in blood cells (12% of expressing cells), skin (19% of expressing embryos, Fig. 3.19C), and developing median fin fold (32% of expressing embryos) of day 2 embryos.

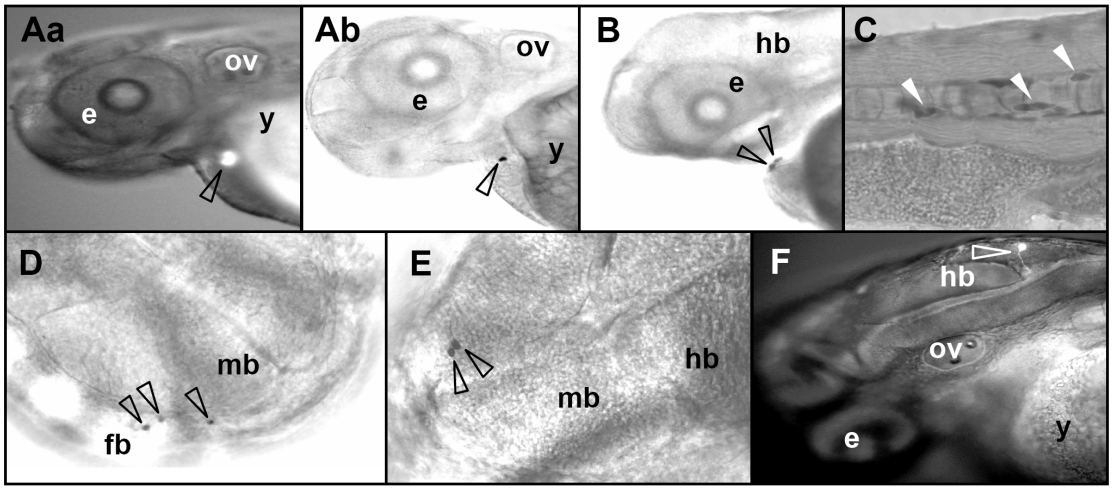


Figure. 3.19. CNE1 mediated reporter expression was more prominent in brain and heart.

GFP expression is shown in fixed tissues after wholemount anti-GFP immunostaining, bright field views (Ab, B, C, D, and E), or GFP fluorescence is shown in live embryos, combining bright field and fluorescence (Aa and F). Arrowheads indicate GFP expressing cells. Presented here are GFP expressing cells (Aa) in cardiac chamber (GFP signal beating with heart) in live embryo (~26-33 hpf); (Ab) GFP expression in the cardiac chamber of an embryo (~48 hpf) shown in fixed tissue following whole mount anti-GFP immunostaining; (B) GFP expression in pericardial muscles at ~ 48 hpf, (C) epidermal cells in the trunk region; (D-F) GFP expressing neurons in forebrain, midbrain, and hindbrain, respectively. e, eye; f, fin; fb, forebrain; h, heart; hb, hindbrain; I, lens; mb, midbrain; ov, otic vesicle; y, yolk.

CNE2

CNE1 and CNE2 are both residing within intron 2 of *GLI3* at its proximal and distal end, respectively. About 60-kb non-coding interval in between these two conserved non-coding elements harbor no additional tetrapod/teleost conservation. Consistent with the physical proximity, the CNE2-mediated GFP expression pattern (Fig. 3.25) was very similar to CNE1 except in the heart chambers where CNE2 injected embryos did not show any GFP signal either on day 2 or day 3 of development.

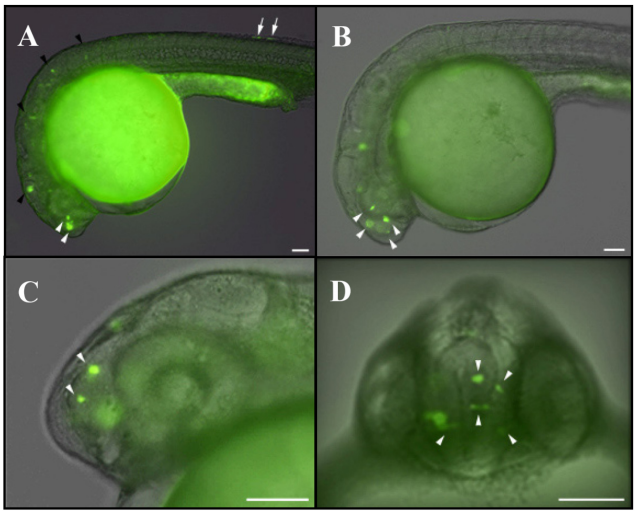


Figure. 3.20. The target site specificity of CNE2 largely overlaps with CNE1.

GFP expression is shown in live embryos, combining bright field and fluorescence. (A) embryo showing expression in the forebrain (white arrowheads), other regions of the CNS (black arrowheads), and in the fin (white arrows); lateral view, anterior to left, dorsal to top (~26-33 hpf). B-D embryo showing expression in the forebrain (white arrowheads), B and C lateral view, anterior to left, dorsal to top; D ventral view. Scale bar = 100 µm.

Most notable activity domains for CNE2 were forebrain (36% of expressing embryos, Fig. 3.20A-D), midbrain (53% of expressing embryos), hindbrain (36% of expressing embryos), and developing median fin fold (23% of expressing embryos Fig. 3.20A).

CNE10

CNE10 directed reporter gene expression most frequently in the eye (54% of expressing embryos), pericardial region (57% of expressing embryos; Fig. 3.21B), and skin cells (48% of expressing embryos). Within the eye, CNE10 mediated reporter expression in the retinal ganglion cells, the photoreceptor layer at the retinal margin, the lens epithelial cell layer, and the lens nuclear region (Fig. 3. 21A & C-E).

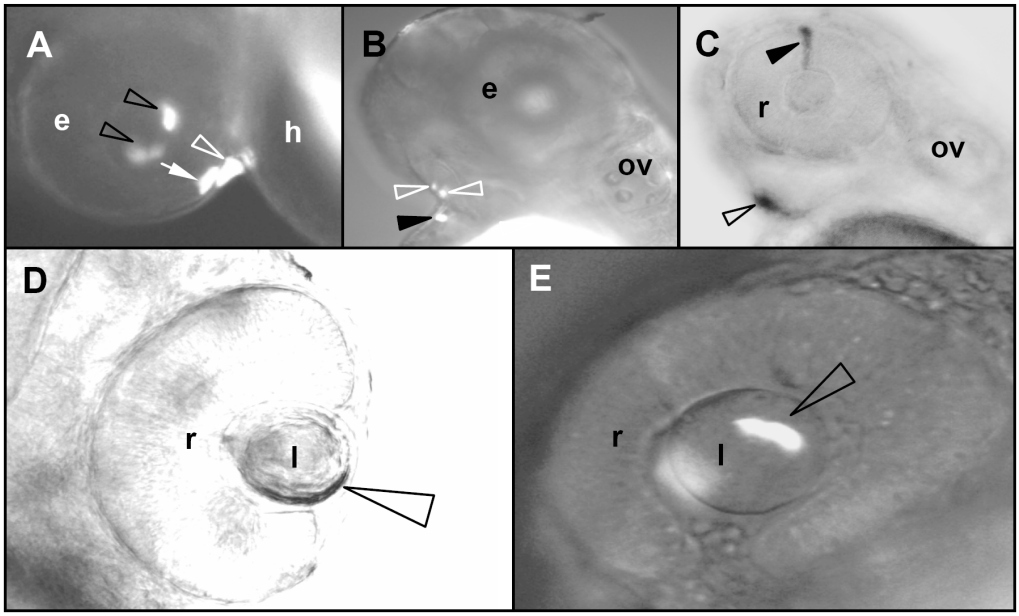


Figure. 3.21. The most frequent domains of CNE10 activity were various subregions of eye and lower jaw.

GFP expression is shown in fixed tissues after wholemount anti-GFP immunostaining, bright field views (C & D), or GFP fluorescence is shown in live embryos, combining bright field and fluorescence (A, B & E). Black, white, and open arrowheads indicate GFP expressing cells. Presented here are GFP expressing cells within CNE10 injected embryos. (A) open black arrowheads indicates reporter expression at the margin of lens placode, whereas a white arrow and a white open arrowhead depict the GFP expressing cells at the margin of retina and at an intermediate dorsoventral position within the first pharyngeal arch (mandibular arch) of day 2 embryo (~26-33 hpf), respectively. (B) white and black arrowheads show GFP expression within lower jaw primordia and pericardial regions, respectively, of a live day 3 (~48 hpf) embryo. (C) Wholemount anti-GFP immunostaining revealed GFP expression within lower jaw primordia (open arrowhead) and a retinal cell (black arrowhead) of a day 3 embryo. (D) An open arrowhead indicates GFP expression within the lens epithelial cell layer. (F) GFP expression in lens nuclear region of day 2 embryo. e, eye; f, fin; h, heart; hb, hindbrain; l, lens; ov; otic vesicle; r, retina; y, yolk.

In addition to eye, pericardial region, and skin, CNE10 also mediated GFP expression uniquely in the lower jaw primordia or first pharyngeal arch (mandibular arch) region in a significant proportion of expressing embryos (24% Fig. 3. 21A-C).

CNE6 & CNE7

CNE6 mediated GFP expression most prominently in the dorsal spinal cord neurons (21% of expressing embryos, Fig. 3.22C) just posterior to the hindbrain/spinal cord boundary. GFP expression was also observed in hindbrain neurons (10% of expressing embryos, Fig. 3.22B), immediately flanking the hindbrain/spinal cord boundary. In addition to hindbrain/spinal cord boundary regions, CNE6 also directed reporter expression in blood cells (17% of expressing embryos, Fig. 3.22D), and muscle fibers (10% of expressing embryos, Fig. 3.22A).

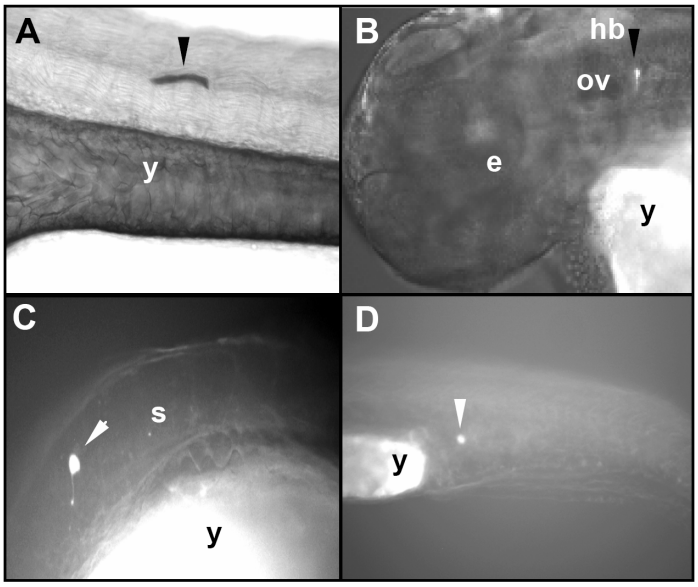


Figure. 3.22. Within zebrafish neural tube the CNE6 activity was restricted to the neuronal populations at the junction of hindbrain and spinal cord.

GFP expression is shown in fixed tissues after wholemount anti-GFP immunostaining, bright field views (A), or GFP fluorescence is shown in live embryos, combining bright field and fluorescence (B-D). Black and white arrowheads indicate GFP expressing cells. Presented here are GFP expressing cells within CNE6 injected embryos. (A) arrowhead indicates the GFP expressing muscle fiber in the trunk of day 3 (~ 48 hpf) embryo. (B) GFP expressing neuron in the hindbrain (~48 hpf) with ventrally extending axon. (C) Just posterior to hindbrain/spinal cord boundary a spinal cord neuron with axon extending ventrally express GFP in day 2 embryo (~26-33 hpf). (D) Blood cell expressing GFP in the region of blood islands. e, eye; hb, hindbrain; ov, otic vesicle; r, retina; s, spinal cord; y, yolk.

CNE7 did induce reporter gene expression in different regions of day 2 embryos (Fig. 3.25), however, the activity was not specific with respect to a particular tissue/region of the embryo.

CNE11

CNE11 activity on day 2 of development was confined to skin cells (64% of expressing embryos), muscle fibers (30%), and heart chambers (30%). In contrast to other elements which drove expression mainly on day 2 (~26-33 hpf) of development, CNE11 strongly enhanced reporter expression also on day 3 (50-54 hpf) of development (Fig. 3.25) within heart chambers (55%), skin cells (25%), muscle fibers (12% Fig. 3.23B), and less prominently in the pectoral fins (Fig. 3.23A).

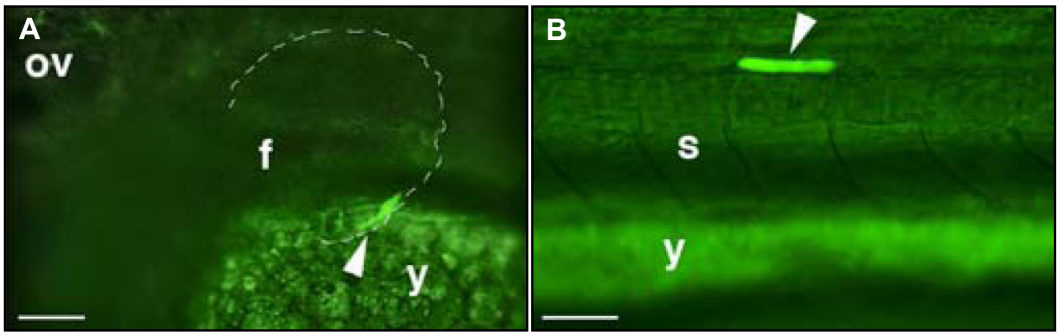


Figure. 3.23. CNE11 governed reporter expression in the pectoral fin bud and a muscle fiber.

GFP expression is shown in live embryos, combining bright field and fluorescence. Black, white and open arrowheads indicate GFP expressing cells. Presented here are GFP expressing cells within CNE11 injected embryos (A) at the margin of pectoral fin of day 3 embryo (~50-54 hpf) and (B) a muscle fiber in the trunk region of day 3 embryo. f, fin; ov, otic vesicle; s, spinal cord; y, yolk.

CNE9

The most prominent GFP expression domain for CNE9 injected embryos was notochord cells (74% of GFP expressing embryos, Fig. 3.24 C-D). In addition to notochord, CNE9 induced reporter gene expression in spinal cord (14% of expressing embryos, Fig. 3.24 A-B), forebrain (11% of expressing embryos), hindbrain (11% of expressing embryos), skin cells (20% of expressing embryos), fin (14% of expressing embryos, Fig. 3.24E) and muscle fibers (11% of expressing embryos, Fig. 3.24F).

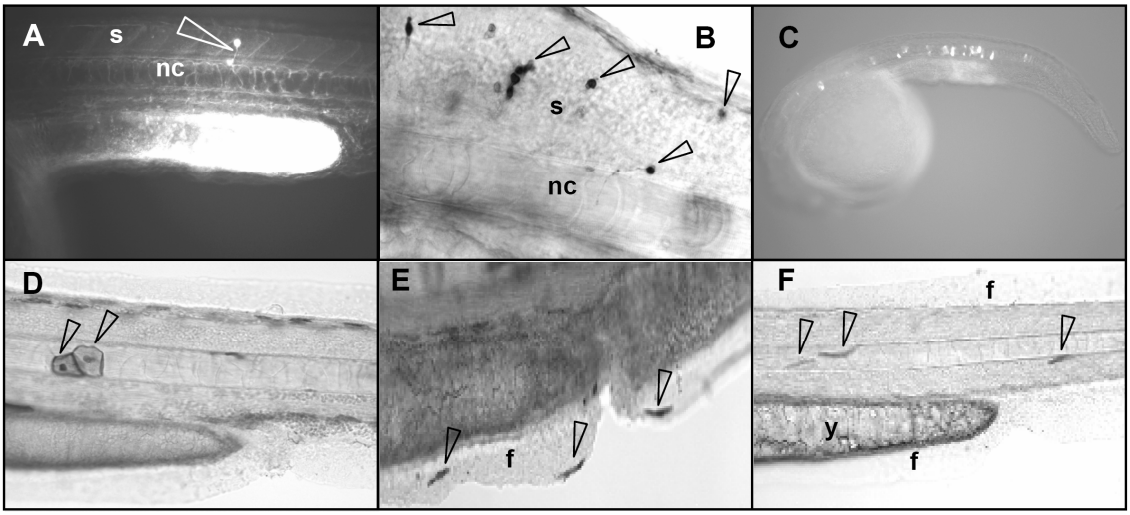


Figure. 3.24. In zebrafish embryos the most obvious domain of CNE9 function was notochord.

GFP expression is shown in fixed tissues after wholemount anti-GFP immunostaining, bright field views (B, D & E), or GFP fluorescence is shown in live embryos, combining bright field and fluorescence (A, C & F). Black, white, and open arrowheads indicate GFP expressing cells. Presented here are GFP expressing cells within CNE9 injected embryos. (A) An arrowhead indicates the GFP expressing spinal cord neuron in a live day 2 (~26-33 hpf) embryo with an axon extending ventrally towards notochord. (B) GFP expressing spinal cord neurons in the anterior trunk region of a day 3 embryo (~48 hpf) shown in fixed tissue following whole mount anti-GFP immunostaining. (C) GFP expression throughout the notochord in a live day 2 embryo (~26-33 hpf). (D) In the trunk region two notochord cells express GFP as shown in fixed tissue following whole mount anti-GFP immunostaining. (E) GFP expressing cells in the developing caudal and ventral fin fold regions. (F) GFP expressing muscle fibers in the trunk region of day-3 embryo. e, eye; f, fin; nc, notochord; s, spinal cord; y, yolk. Anterior to left, dorsal to top.

3.6 Generalized scheme of GFP expression domains in zebrafish embryos at 26-33 hpf or 50-54 hpf

For each CNE hundreds (average 200-embryos/CNE) of embryos were microinjected, and subsequently the expression data from all positive embryos for each element was overlaid onto a schematic diagram, to give an overall impression about the activity domains associated with each *GLI3* associated enhancer element (Fig. 3.25).

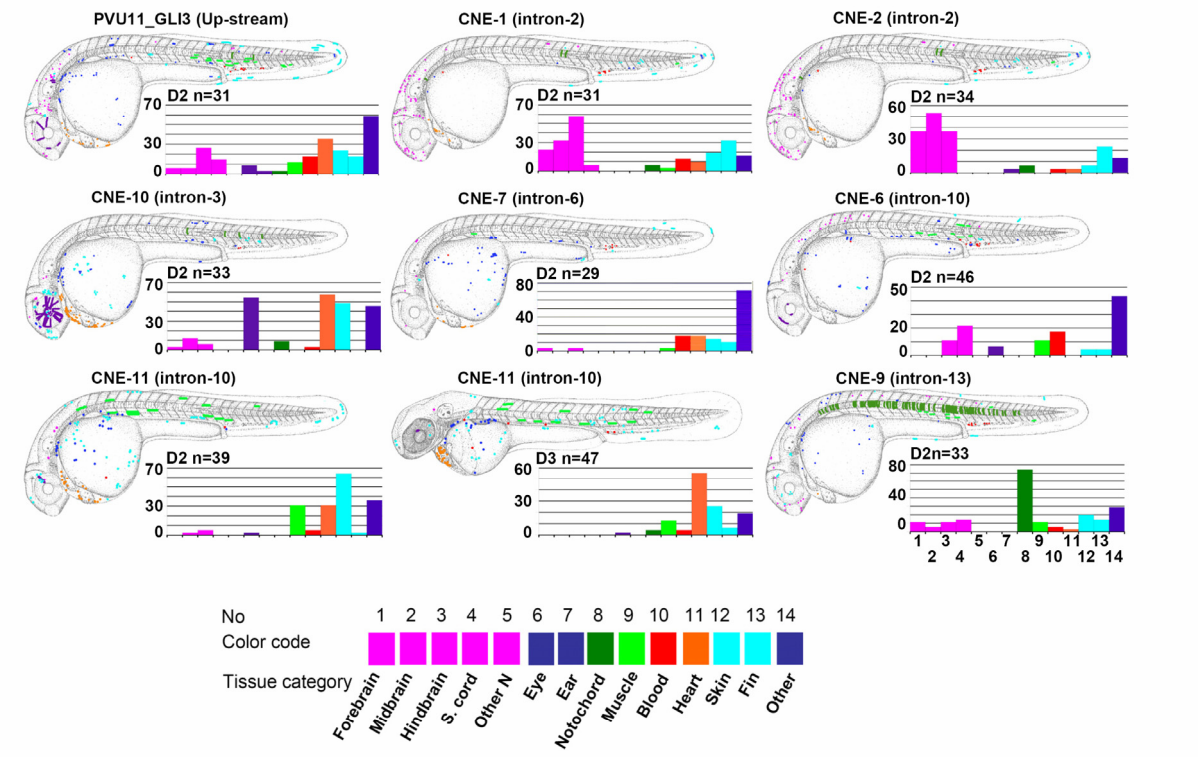


Figure. 3.25. Sites of GFP expression induced by *GLI3*-associated CNEs in zebrafish embryos

Sites of GFP signals recorded in zebrafish embryos transiently transfected with a construct, in which the reporter gene was induced by individual *GLI3*-associated CNEs (indicated by name and location in a *GLI3*-intron), are depicted in schematic representations of day two (24-33 hpf, D2) or day 3 (D3) embryos. The total number of positive embryos per CNE (n) is indicated for each of the constructs. Categories of cell type that were positive for a given element are color 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 color code as the schematics for each cell type.

3.7 One out of four conserved non-coding elements from the intronic region of *GLI3* showed weak enhancer activity in the chicken limb bud

Four conserved non-coding elements from the intronic region of *GLI3* were tested in collaboration with the laboratory of Professor Cheryll Tickle, Dundee, UK, for enhancer activity in the chicken limb. The putative enhancers CNE1, 9, 10 ,11 were cloned into a GFP reporter construct under the control of a β-globin promoter. The chicken limb was co-electroporated at stage HH 19/20 *in ovo* with each of these constructs and an RFP reporter to control for electroporation efficiency. The embryos were assessed for RFP and GFP expression 48 hours following electroporation when they had reached approximately stage HH 26. At this stage, Gli3 has been reported to be expressed at the distal anterior edge of the limb (Schweitzer et al. 2000) (see also Fig. 3.26d''). Fig. 3.26 shows the results of the electroporation experiments, which are also detailed in Table 3.2. The upper row (Fig. 3.26a-d) shows RFP expression in the limb bud indicating the extent of electroporation. The middle row (Fig. 3.26 a'-d') shows GFP expression in the same limb bud. CNEs 1, 9 and 10 gave no GFP expression (Fig. 3.26a'-c', Table 3.2) despite RFP being expressed throughout the limb (Fig. 3.26 a-c). CNE11 had weak GFP expression in 2/6 cases (Fig. 3.26 d,d' arrows, Table 3.2), indicating slight enhancer activity in the limb bud. The distribution of Gli3 mRNA in the limb at stage HH26 is shown below via *in situ* hybridisation (Fig.3.26 d''). Gli3 was highly expressed distally but also proximally at the posterior margin and, therefore, reporter activity appears to be within the region of the limb expressing Gli3 (Fig. 3.26d'', arrow). At earlier stages, Gli3 is more highly expressed throughout the anterior of the limb bud (Fig. 3.26e'') and, therefore, we might expect that electroporation of the putative enhancer constructs at an earlier stage would provide a better test for enhancer activity. In a second set of experiments we, therefore, electroporated CNE11 into the presumptive limb mesenchyme at stage HH14 and then looked for enhancer activity 48 hours post electroporation at approximately stage HH23. RFP expression was found throughout the anterior region of the limb (Fig. 3.26e), however, no GFP expression was seen in any of the cases examined (Fig. 3.26e'), although the construct had been successfully electroporated into the region of the limb bud, which would be expressing Gli3 (compare RFP expression in Fig.3.26e with Gli3 expression in Fig. 3.26e'').

Table 3.2. Electroporation of the recombinant constructs carrying *GLI3* enhancers in combination with a mouse β-globin promoter / GFP reporter system.

CNEs	Length	GFP expressing embryos HH23	GFP expressing embryos HH26
CNE1	945bp	-	(-) 0/5
CNE9	669bp	-	(-) 0/7
CNE10	1133bp	-	(-) 0/2
CNE11	1185bp	(-) 0/5	(+) 2/6

Note Results are detailed as enhancer activity detected (+) or no enhancer activity (-). The proportions of the total number of embryos electroporated are given.

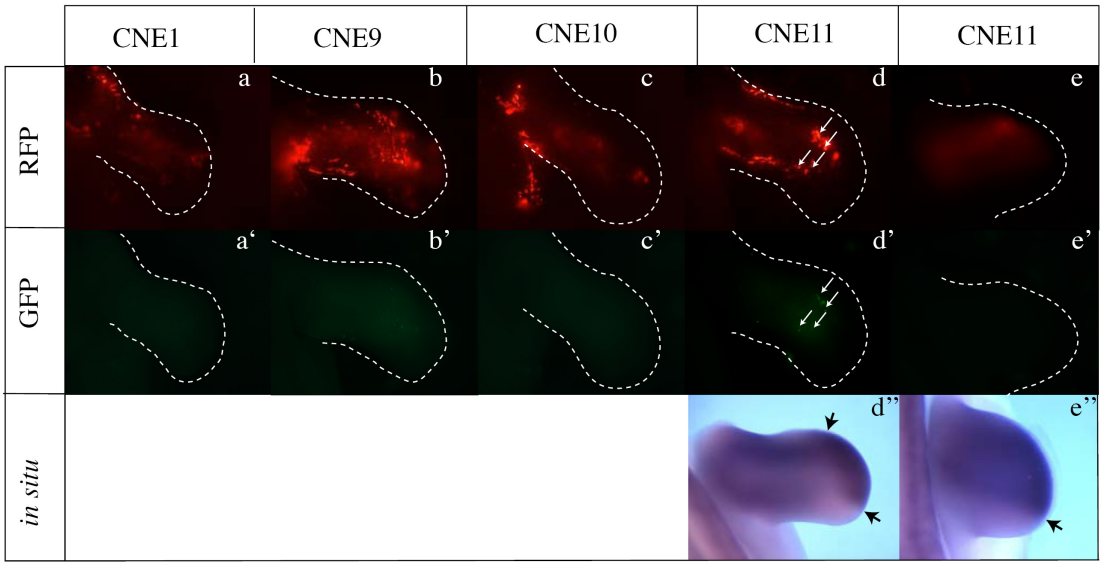


Figure. 3.26. Enhancer analysis in chicken limb buds

a-d) Whole mount of limb buds 48 hours after co-electroporation of enhancer construct and RFP at stage HH20. e) Whole mount of limb bud 48 hours after co-electroporation of CNE11 and RFP at stage HH14. RFP indicates electroporated regions (a-e). Presence of GFP indicates enhancer activity (a'-e'). CNE11 shows enhancer activity when electroporated at HH19/20 (GFP and corresponding RFP are indicated with arrows) but not when electroporated at HH14. CNE 1,9, and 10 do not show any GFP expression. The bottom row (d'' and e'') shows the distribution of Gli3 transcripts via *in situ* hybridisation. At stage HH26 transcripts are seen at the distal tip (arrows 1d'') and more proximally, but absent from the distal posterior region. Transcripts are seen in the anterior of the limb at stage HH23 (between arrows e''). Note that GFP expressing cells in 1d' are found in a region of the limb where Gli3 appears to be expressed.

3.8 Expression of a reporter construct in transgenic mice under the control of CNEs

The subset of *GLI3* associated CNEs that acted as enhancers in cell culture and zebrafish were screened for their ability to drive the expression of a reporter in transgenic mice. In this regard, recombinant constructs have been generated with the vector "p1230", in which a bacterial *lac*Z-reporter gene is placed under the control of a ß-globin minimal promoter. This vector had previously been applied in numerous studies (Lettice et al. 2003), and is known not to be expressed in transgenic mice without additional enhancer elements. Recombinant constructs carrying GLI3_ PvuII, CNE1, 6, 9, 10, and 11 were injected into the pronuclei of fertilized mouse oocytes using standard procedures to generate transgenic mice. After cloning, the vector sequences were separated at appropriate polylinker restriction sites from the enhancer/*lac*Z reporter-insert, and the fragments were prepared and purified for microinjection.

For each of the enhancer elements embryonic mice from stable transgenic lines have been analyzed at different time points of development by whole mount staining and using histological sections to determine the time course and the location of reporter X-Gal signals reflecting expression of the reporter gene. Histological analysis of mouse embryos was performed in collaborations with the lab of Dr. Ansgar Schmidt, Marburg, Germany.

GLI3_PvuII

The human DNA fragment of ~3.5 kb, encompassing part of the *GLI3* promoter, showed a widespread activity in mouse embryos. At embryonic day 9.5 the reporter expression was confined to hindbrain, in E11.5 embryos β-galactosidase staining was also observed in the mesencephalic neural tube and in spinal cord, while at E12.5 the reporter expression, in addition, was present in the both fore- and hindlimb buds (Fig. 3.27A-D). Close histological examination of E12.5 embryos revealed the presence of *lacZ* activity in the dorsolateral aspects of diencephalon and mesencephalon as well as in the basal plate region of the pons. Moreover in the spinal cord the *lacZ* activity was observed in the floorplate and throughout the mantle layer (Fig. 3.27E-H). Histological analysis also revealed β-galactosidase staining in paired somites within the dorsal medial lip of the dermamyotome (arrowheads in Fig. 3.27H), and in the urinary bladder (Fig. 3.27I).

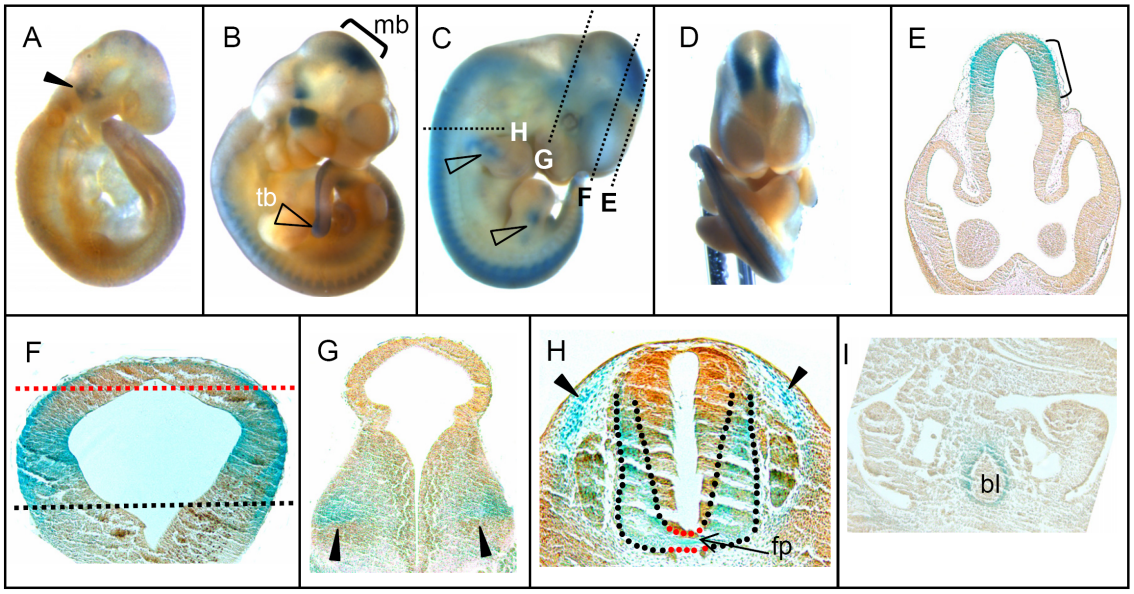


Figure. 3.27. Enhancer analysis of the element GLI3_PvuII in transgenic mouse embryos

Expression of a β-galactosidase reporter gene in embryos from a permanent transgenic line carrying a ~3.46kb DNA fragment encompassing part of the human *GLI3* promoter fragment. Whole mount views of (A) E9.5, (B) E11.5, (C & D) E12.5 embryos. At E9.5, β-galactosidase staining is confined to hindbrain (black arrowhead in A) while in E11.5 embryos (panel B) reporter expression is also seen in the midbrain and spinal cord. In day 12.5 embryos, β-gal activity is observed in the proximal region of fore- and hindlimb buds (open arrowheads in panel C). (D) Frontal view of E12.5 shows reporter expression in the lateral aspects of mesencephalon. (E-I) transverse sections of E12.5 embryos at the levels shown with dotted lines in panel C. β-galactosidase is restricted to the lateral walls of the diencepalon (shown in panel E with bracket symbol) and mesencephalon (red and black dotted lines indicates the dorsal and ventral boundary of reporter expression). In panel G the black arrowheads show staining in the basal plate region of pons. In the neural tube, the reporter expression was observed in the floorplate (demarcated by red dots) and throughout the mantle layer (demarcated with black dots); black arrowheads in panel H indicate staining in the paired somites. (I) *lacZ* expression in the urinary bladder. tb, tail bud; mb, midbrain; fp, floorplate; bl, urinary bladder.

CNE1

Two independent permanent transgenic lines carrying the 935 bp CNE1 show similar β-gal activity. At E9.5 strong *lacZ* expression was observed throughout the dorsal aspects of rostral-caudal axis of neural tube. In addition staining was also detected in the first (maxillary and mandibular components are shown by white and black arrowheads, respectively) and second branchial arch regions (Fig. 3.28A). At E11.5 the expression is maintained in the dorsal brain and spinal cord (Fig. 3.28A-C), whereas within branchial arches no reporter activity was detected in the mandibular components of first arch (Fig. 3.28B-E). In day E11.5 embryos, the reporter expression was also observed in hypaxial buds of the thoracic somites, proximal muscle masses in the forelimb bud, dorsal root ganglion, and in the facial mesenchyme (Fig. 3.28 B & E). At E12.5 the reporter expression pattern was largely similar to that of E11.5 (compare 3.28B with F, and C with G) except that a strong

reporter activity was seen in the cerebellum (Fig. 3.28G) and in the spinal nerves innervating dorsolateral trunk region and forelimbs (Fig. 3.28I). In addition, within the head mesenchyme, the β-gal activity extended more rostrally (white arrowhead in panel 3.28H, also compare 3.28D with H).

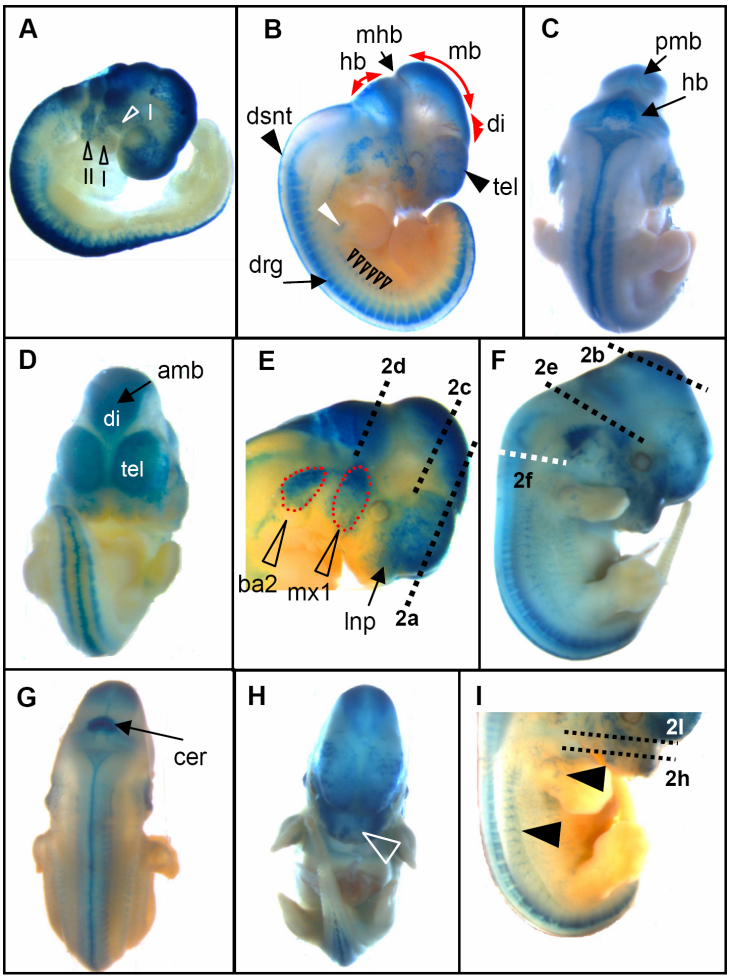


Figure. 3.28. CNE1 governed *lacZ* expression, primarily in the neural tube and facial regions of the transgenic mouse embryos

Expression of a β-galactosidase reporter gene in transgenic mouse embryos. Whole mount views of E9.5 (A), E11.5 (B-E), and E12.5 (F-I) embryos. At E9.5 (A), CNE1 mediated reporter expression was present throughout the dorsal aspects of neural tube, in addition the β-gal staning was also detected in the maxillary/mandibular component of first branchial arch (white and black arrowheads in panel A, labeled as I) and in the second arch (black arrowhead in panel A, labeled as II). At E11.5 (B-E), the expression is maintained in dorsal telecephalon, diencephalon, midbrain, hindbrain, dorsal spinal cord (compare panel B with A), maxillary component of first arch, and in the second branchial arch (E). In addition, in day 11.5 embryos the reporter signals were also detected in hypaxial buds of the somites (open arrow heads in B) and within proximal muscle masses in the forelimb bud (white arrowhead in B). (F-I) at E12.5 the reporter expression pattern was largely similar to that of E11.5 (compare B with F and C with G), except that strong reporter activity was seen in cerebellum (G) and the spinal nerves inervating the dorsolateral trunk region and forelimb (black arrowheads in I). In addition, within the facial region, the β-gal activity extended more rostrally (white open arrowhead in H, also compare D with H). A, B, E, F and I lateral views, C and G dorsal views, D and H ventral views. tel, telencephalon; di, diencephalon; mb, midbrain; pmb, posterior midbrain; amb, anterior midbrain; mhb, midbrain-hindbrain boundry; hb, hindbrain; dsnt, dorsal neural tube; drg, dorsal root ganglion; mx1, maxillary component of first branchial arch; ba2, second branchial arch; lnp, lateral nasal process; cer, cerebellum; drg, dorsal root ganglia.

In histological analysis of whole-mount embryos, it appeared that within the forebrain the *lacZ* expression is present throughout the wall of telencephalon, however, signals were more widespread medially, whereas in the diencephalon CNE1 directed *lacZ* expression specifically within its dorsal (prospective epithalamus) and ventral aspects (prospective hypothalamus) (Fig.3.29A). At the level of midbrain, transgene expression was present broadly in the dorsal (dorsal midline) and dorso-lateral portion of alar plate, and was confined to marginal layer of basal plate, excluding the ventral mid-line (floorplate) (Fig. 3.29B, C). In the rostral part of the hindbrain the reporter expression driven by CNE1 was detected dorsally throughout the wall of cerebellum, whereas ventrally it was confined to neuronal cell populations within tegmentum of pons (Fig. 3.29C). In the caudal part of hindbrain, in addition to the roofplate, the *lacZ* expression was detected in two discrete neuronal cell populations within the medulla oblongata (Fig. 3.29E). Further histological examination demonstrated that transgene expression was present in the regions occupied by four different classes of ventral interneurons (V0-V3) of the spinal cord (Fig. 3.29F). Additionally, the histological sections through the facial regions revealed *lacZ* activity in the medial and lateral nasal processes, precartilage primordium of nasal capsule, Meckel's cartilage, lateral palatine process, and in the dental lamina (Fig. 3.29G-I).

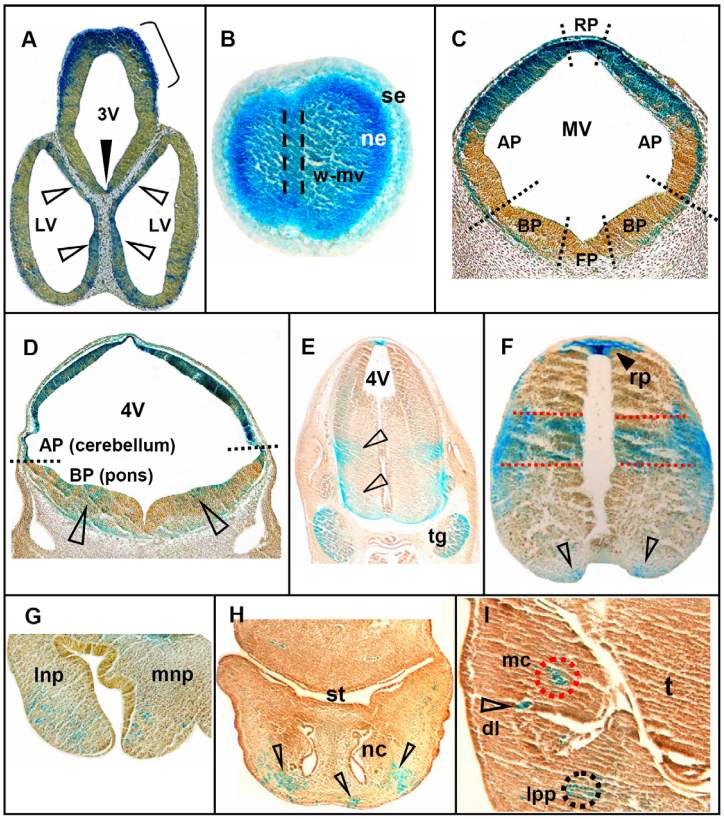


Figure. 3.29. Histological inspection of embryos stably carrying a reporter transgene activated by CNE1

Histological analysis of E11.5 (A, C, D and G) and E12.5 (B, E, F H and I) embryos from permanent transgenic lines carrying CNE1, following β-galactosidase staining. Transverse scetions of transgenic embryos were obtained at the levels shown with dotted lines in the panels E, F and I of Fig. 3.28. Within the lateral ventricles of forebrain the reporter expression was widespread medially (open arrowheads in panel A) and was confined to scattered cell pouplations laterally, whereas in the diencephalon the reporter expression is evident in the dorsal (shown with bracket symbol in A) and ventral (black arrowhead in panel A) aspects. B: transverse section through the top of midbrain shows strong reporter expression signals in the neuroepithelium surrounding the dorsal (dorsal midline, marked with dotted lines in panel B) and lateral aspects of mesencephalic vesicle. C: in the roofplate and dorsolateral portion of alar column of midbrain the *lacZ* expression is present in all three distinct layers (marginal, mantle, and ependymal) of neuroepithelium, whereas in the medial portion of alar plate/entire basal plate regions of midbrain the staining is detected only in the marginal layer. Note, reporter expression is excluded from the floorplate of midbrain. D: reporter expression is widespread in the alar plate region (presumptive cerebellum) of metencephalon, whereas in the basal plate region the reporter activity is confined to neuronal cell populations in the tegmentum of pons (open arrowheads). In panel E the open arrowheads indicate β-galactosidase staining in the two discrete neuronal cell populations within the caudal part of medulla oblongata (myelencephalon). Staining is also evident in trigeminal (V) ganglion. F: Within spinal cord, the *lacZ expression* is confined to roofplate (black arrowhead) intermediate neurons (the staining in the area, encompassing V0, V1 and V2 interneurons is bounded by red dotted lines) and V3 interneurons (indicated by open arrowheads). G: At E11.5, the reporter expression is evident in the lateral nasal process and medial nasal process. H: open arrowheads indicate β-galactosidase staining in the precartilage primordium of nasal capsule of E12.5 embryo. I: red and black dotted circles highlight reporter expression signals in the precartilage primordium of Meckel's cartilage and lateral palatine process, respectively, whereas in the same panel the open arrowhead shows β-galactosidase staining within dental lamina (tooth primordium). 3V, third ventricle; LV, lateral ventricle; se, surface ectoderm; ne, neuroepithelium; w-mv, wall of mesencephalic vesicle; RP, roofplate; MV, mesencephalic vesicle; AP, alar plate; BP, basal plate; FP, floorplate; 4V, fourth ventricle; tg, trigeminal ganglion; lnp, lateral nasal process; mnp, medial nasal process; st, stomodaeum; nc, nasal cavity; mc, Meckel's cartilage; llp, lateral palatine process; dl, dental lamina; t. tongue;

CNE2

Intra-*GLI3* conserved non-coding region 2, is among those 481 human genomic fragments, that show 100% identity over ≥ 200 bp in human, mouse and rat comparison, and are distinctively known as "Utraconserved Elements" (Bejerano et al. 2004). In the transgenic mouse assay, the primary activity domains of CNE2 were forebrain and facial regions. The in depth functional analysis of this interval in mice is presented elsewhere (Paparidis et al. 2007).

CNE10

The whole mount view of the E11.5 embryos from a stable transgenic line carrying CNE10 revealed transgene expression in the heart (Fig.3.30A), whereas at E12.5 the CNE10 mediated lacZ expression was observed in the eye, mammary placodes, and in the external genitalia region (Fig. 3.30B-D). Furthermore, the β-galactosidase staining of the dissected E12.5 embryos revealed strong reporter expression in the urogenital system, stomach and pancreatic bud (Fig. 3.30 E-F). Additionally, detailed histological examination of E11.5 embryos demonstrated the CNE10 mediated lacZ expression in visceral organs, including ventricular chamber of heart, pharynx, oesophagus, tracheal duct, liver, gall bladder, wall of stomach, dorsal pancreatic primordium, lumen of duodenum, proximal loop of midgut, mesonephric vesicles and mesonephric (Wolffian) duct (Fig. 3.30J-P). β-

galactosidase was also seen along the roof of myelencephalon, in the ventral region of optic stalk, sensory layer of retina, and in Rathke's pouch (Fig. 3.30G-I and L).

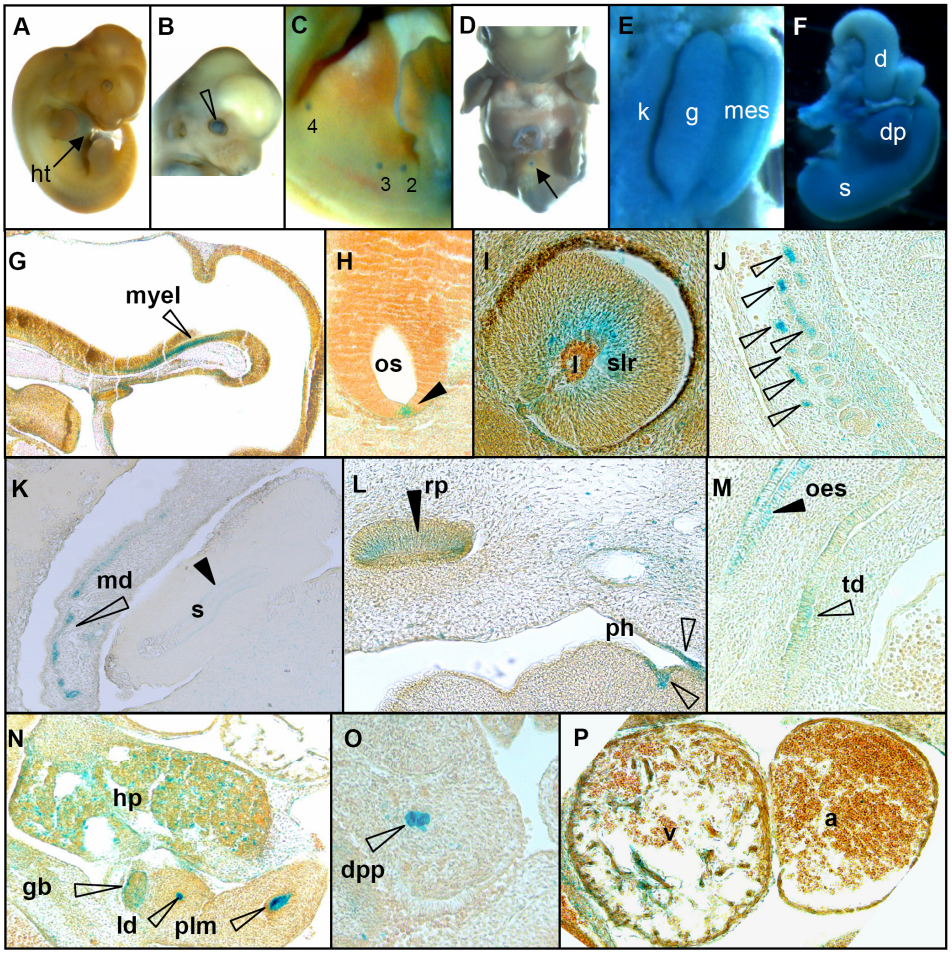


Figure. 3.30. Histological analysis of CNE10 harboring embryos

Whole mount views of E11.5 (A) and E12.5 (B, C and D) embryos carrying a reporter construct encompassing the 1133 bp CNE10. In E11.5 embryos the CNE10 mediates *lacZ* expression in heart (arrow in A) and other visceral organs. In E12.5 embryos *lacZ* expression was observed in eye (B) in mammary placodes (numbered as 2, 3, and 4). Note, mammary placodes 1 and 5 are covered by forelimb and hindlimb and are not visible (C). An arrow indicates staining in the external genitalia region (D). In panels E and F, staining in urogenital system (E) and stomach and pancreatic bud from dissected E12.5 embryo. Note, the staining was performed after dissection. Sagittal sections of the transgenic E11.5 embryos (G-P). Sectioning revealed *lacZ* expression: (G) along the roof of myelencephalon; (H) in the ventral region of optic stalk; (I) sensory layer of the retina of eye. (J) Arrowheads indicate β-galactosidase staining in the mesonephric vesicles. (K) Open and black arrowheads display reporter expression in the mesonephric duct and along the wall of stomach (weak staining), respectively. (L) Staining in Rathke's pouch and pharynx. (M) Black and open arrowheads show reporter expression in the oesophagus and tracheal duct. (N) Reporter expression in the hepatic primordium (liver), gall bladder, lumen of duodenum and proximal loop of midgut. (O) Strong expression is seen in the pancreas. (P) β-galactosidase is present in the ventricular chamber of the heart but not in the atrial chamber. ht, heart; v, ventricular chamber of heart; a, atrial chamber; l, live; s, stomach; myel, myelencephalon; os, optic stalk; l, lense; slr, sensory layer of retina; md, mesonephric duct; rp, Rathke's pouch; p, pharynx; hp, hepatic primordium; gb, gall bladder; ld, lumen of duodenum; plm, proximal loop of midgut; dpp, dorsal pancreatic primordium; oes, oesophagus; td, tracheal duct .

CNE6

The 862 bp CNE6 directs β-galactosidase reporter expression primarily within forelimb and hindlimb buds (Fig. 3.31A, B &D). At embryonic day 9.5 and 10.5 strong β-galactosidase staining is seen throughout the limb mesenchyme, except in the posterior region (Fig. 3.31G & H), which corresponds to the zone of polarizing activity (ZPA), and expresses *Shh* to organize anterior-posterior patterning during mouse limb development. In E11.5 forelimb, *lacZ* expression is seen in the proximal and distal regions of limbs, but absent from distal posterior region (Fig. 3.31I). At E12.5, CNE6 governed transgene expression prominently within the anterior half of the handplate, including prospective digits 1-4, and the interdigital mesenchyme between them, whereas within the posterior part the *lacZ* expression was undetectable (Fig. 3.31J). In addition to limbs, the CNE6 directed reporter expression was also evident in rostroventral telencephalon in E9.5 and 10.5 embryos (open arrowhead in Fig. 3.31A), and later, by E11.5, in the head mesenchyme in the region of nasal process (Fig. 3.31C & E).

Close histological examination of E12.5 forelimb revealed that, in addition to non-condensing mesenchyme, the CNE6 mediated transgene expression also occurred in the precartilage condensations of mesenchyme specifically within digital rays (Fig. 3.31K). Additionally, histological analysis confirmed the whole mount observations, that within non-condensing mesenchyme the CNE6 activity is largely confined to cellular populations within anterior half of the handplate (compare reporter expression pattern shown in Fig. 3.31K and J). Histological examination of the facial region indicated the *lacZ* activity, at E12.5 specifically within the precartilage primordium of nasal septum (Fig. 3.31F).

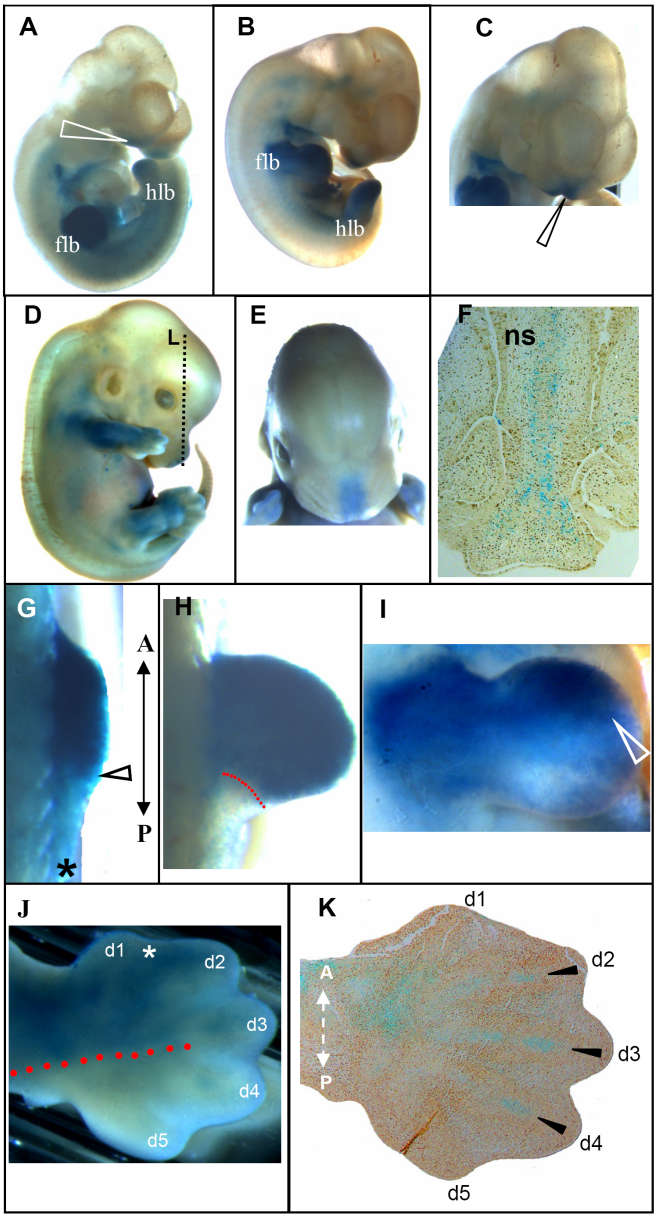


Figure. 3.31. CNE6 induced reporter activity at different time points of developing mouse embryos

Expression of a β-galactosidase reporter in embryos from a permanent transgenic line carrying the 865 bp CNE6. In Panel A-E the whole mount views of E10.5 (A) 11.5 (B & C) and 12.5 (D &E) embryos showing staining in the limbs (A, B and D), rostroventral forebrain (white arrowhead in A) and in the head mesenchyme in the region of nasal process (C and E). (F) Transverse section of the E12.5 embryo at the level shown with black dotted line in panel D, revealed transgene expression specifically within precartilage primordium of nasal septum. In panel G-J, the limb buds are focused to show precisely regions of transgene activation at different time points. In E9.5 (G) and E10.5 (H) embryos, strong β-galactosidase staining is seen throughout limb buds except in the posterior margin, i.e. the zone of polarizing activity (ZPA). Open arrowhead in panel G and I, and red dots in panel H and J demarcate the posterior border of reporter activity, while the asterisk symbol in panel G shows the posterior margin of the forelimb bud. At E11.5 (I), *lacZ* transcripts are seen in the proximal and distal regions of the limb bud, but absent from distal posterior region. Open arrowhead indicates the distal posterior margin of reporter expression. In E12.5 limb buds (J), the transgene activation occurred within anterior half of the hand plate. The asterisk symbol in the E12.5 hand shows strong reporter activity within interdigital mesenchyme between digits 1 and 2. (K) Longitudinal sections through the E12.5 forelimb revealed *lacZ* expression in the precartilage condensations of mesenchyme within digital rays (prospective phalangeal cartilage). Note, at E12.5 in the wirst/hand region, the staining is largely restricted to the non-condensing mesenchyme within anterior half but excluded from the posterior half. flb, forelimb bud; hlb, hindlimb bud; d, digit; ns, nasal septum.

CNE11

For the 1185 bp CNE11, a similar reporter expression pattern was observed with two different permanent transgenic lines. In E11.5 embryos, CNE11 mediated reporter activity was restricted to the facial nerve (VII) (Fig. 3.32A & E). At embryonic day 12.5, in addition to the facial region, reporter expression was also detected within proximal (stylopod) and intermediate (zeugopod) elements of both forelimb and hindlimb buds (Fig. 3.32B) as well as in the lateral plate mesenchyme of the flank at the level of forelimbs (Fig. 3.32C). However, up-to this time point CNE11 did not induce transgene

expression in the handplate region (autopod) (Fig. 3.32F). At E13.5, the reporter activity was maintained within facial region, whereas in the limbs the reporter activity was extended to the proximal region of the handplate within digit arch region (wrist/ankel and hand/foot). Nonetheless, the *lacZ* expression was excluded from the digital rays and digital interzones (Fig. 3.32D & G).

Histological analysis of E13.5 limbs showed that, the CNE11 induced reporter activity is specifically restricted to the chondrogenic mesenchymal condensations of limb elements, whereas the *lacZ* expression was excluded from the encasing non-condensing mesenchyme (Fig. 3.32H & I).

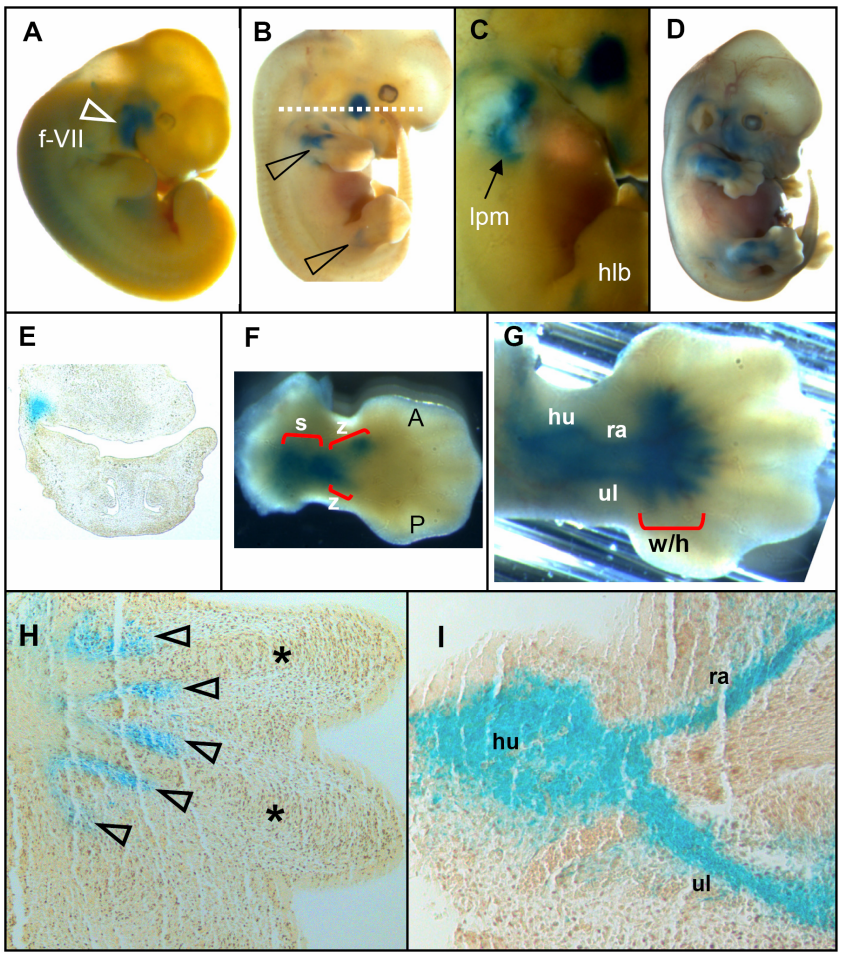


Figure. 3.32. In transgenic mouse embryos carrying CNE11, the *lacZ* expression occurred specifically in the chondrogenic mesenchymal condensations of limb elements, excluding the digits.

β-galactosidase activity in embryos from permanent transgenic lines carrying the 1185 bp CNE11. (A-F) Whole mount views of E11.5 (A), E12.5 (B-C), and E13.5 (D) embryos. A similar expression pattern was repeated with two different lines. In E11.5 embryos (A), CNE11 mediates *lacZ* expression in facial (VII) nerve, whereas at day 12.5 reporter expression was also observed in both fore and hindlimb buds (open arrowheads in B) and in the lateral plate mesoderm at the level of limb buds (C; note forelimb bud was removed to show this pattern). In E13.5 embryos, the lacZ expression was retained in the limbs and in the fascial region (D). In panel E, the transverse section of the E12.5 embryos (at the level shown with white bar in B) shows staining in the facial (VII) nerve. F-G magnified view of E12.5 and E13.5 forelimbs, respectively. In E12.5 forlimb (F), β-gal activity can be seen in the presumtive proximal (stylopod) and intermediate (zeugopod) limb elements, whereas at E13.5 the reporter activity was extended upto the wrist/ankel and hand/foot elements but not to the digits (in panel G, only the forelimb is shown). Longitudinal sections through the lower (H; autopod) and upper extremeties (I; zeugopod/stylopod) of E13.5 forelimb showing *lacZ* expression in metatarsrsal (shown with open arrowheads in panel H), humerus, radius and ulna (panel I) cartilages. Note, no expression is seen in the cartilages of digits (indicated by asterik symbol in panel H). f-VII, facial (VII) nerve; lpm, lateral plate mesoderm; hlb, hindlimb bud; s, stylopod; z, zeugopod; A, anterior; P posterior; hu, humerus ; ra, radius; ul, ulna; w/h, wrist and hand elements.

CNE9

At E10.5, the CNE9 activated transgene expression within anterior midbrain, ventral hindbrain and ventral spinal cord, down to the level of the inter-limb region (Fig. 3.33A). At E11.5 the reporter expression was detected in the dorso-lateral aspects of anterior and posterior midbrain regions, whereas the reporter activity was maintained in the ventral portions of hindbrain and spinal cord up-to the level of forelimb region (Fig. 3.33B). Furthermore *β*-galactosidase staining was also observed in the inter-limb somites in E11.5 embryos (Fig. 3.33F). At E12.5 the CNE9 governed transgene expression was also detected in the medial ganglionic eminence (Fig. 3.33D).

The histological examination of E11.5 and 12.5 embryos indicated that within midbrain *lacZ* expression was widespread ventrally, confined to marginal layer of dorso-lateral aspects, and was excluded from the dorsal midline (Fig. 3.33G-I). In the spinal cord, *β*-galactosidase staining is observed within motor neurons (MN) flanking the floor plate (Fig. 3.33J). Further histological analysis revealed *lacZ* activity in the ventral-lateral lip of dermomyotome within the inter-limb somites (Fig. 3.33K).

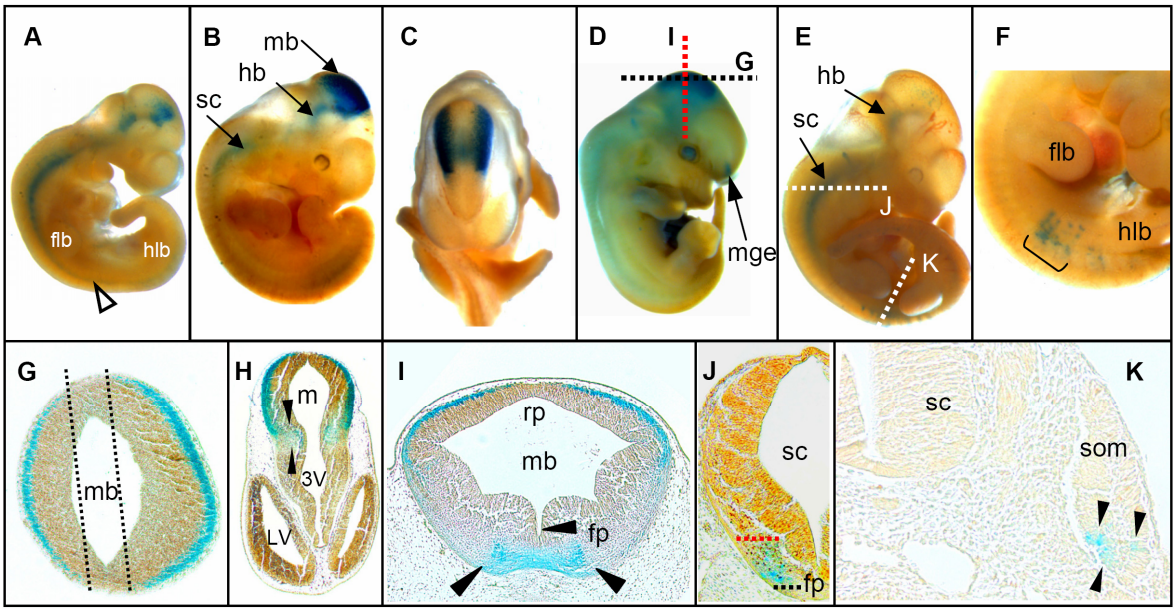


Figure. 3.33. CNE9 induced reporter activity in mouse neural tube was non-redundant to CNE1

Expression of a β-galactosidase reporter gene in embryos from permanent transgenic lines carrying the 669 bp CNE9. Whole mount views of E10.5 (A), E11.5 (B-C, E-F), and E12.5 (D) from two different stable transgenic lines. Stable line 1: A-D; stable line 2: E-F. (G-H) Transverse sections of the embryos at the level shown with dotted lines in panel D and E, stable line 1: G-I (E12.5); stable line 2: J and K (E11.5). (A) At E10.5, within the brain region, weak lacZ expression was restricted to latero-ventral aspects of the anterior midbrain and ventral hindbrain, while in the ventral spinal cord the reporter expression was seen down to the level of the interlimb region (open black arrow head). (B) Arrows depict reporter expression in the midbrain, ventral hindbrain, and ventral spinal cord of E11.5 embryo. (C) Reporter expression is seen at the dorsolateral aspects of midbrain, while absent from dorsal midline of midbrain; (D) at E12.5 β-galactosidase staining was also observed in medial ganglionic eminence (arrow). (E) E11.5 embryos from stable line 2 show reporter expression in the ventral hindbrain and upper cervical region of ventral spinal cord (F), and also in the interlimb (indicated by bracket symbol) somites. (G) dotted lines indicate the exclusion of reporter expression from dorsal mesencephalic midline, while the β-galactosidase staining is present in the dorso-lateral mantle layer of midbrain; (H) at the level of lateral ventricles, the reporter expression within the midbrain expands to mantle and ependymal layer rostrally (arrowheads) but is limited to mantle layer caudally. (I) β-galactosidase staining is observed in ventral midline of caudal midbrain (presumptive dopaminergic neurons). The arrowheads underneath the staining indicate the lateral limit of *lacZ* expression in the ventral midline; (J) in the upper cervical region of spinal cord, β-galactosidase expression within motor neurons (MN) flanking the floor plate, red and black dotted bars demarcate the dorsal and ventral boundaries of reporter expression, respectively; (K) reporter activity within ventral-lateral lip of interlimb somite (arrowheads) . Mb, midbrain; 3V, third ventricle; LV, lateral ventricle; rp, roof plate; fp, floor plate; sc, spinal cord; som, somite; hb, hindbrain; fb, forelimb bud; hb, hindlimb bud.

3.9 Evolution of GLI sequences in vertebrates

3.9.1 Estimation of sequence divergence among species

The pattern of nucleotide substitutions was compared at both silent (synonymous) and non-silent (non-synonymous) sites among GLI orthologs within and between the fish and tetrapod lineages to estimate via the level of sequence divergence the influence of selection at various phylogenetic distances. Selection was measured in terms of the difference in the rate of non-synonymous substitutions (Ka) to the rate of synonymous substitutions (Ks). If Ka and Ks values are not significantly different from each other (Ka = Ks), i.e. if the amino-acid replacement substitutions are observed at the same rate as silent substitutions, then few or no amino-acid replacement substitutions have been eliminated. This indicates that genes are under few or no selective constraints and thus evolving neutrally. A gene pair is said to be under negative selection, if the Ka value is significantly lower than Ks (Ka < Ks), i.e. if non-silent substitutions have been purged by natural selection. The smaller the value of Ka compared to Ks, the larger the number of eliminated substitutions. The converse scenario, where the Ka value is significantly greater than Ks (Ka > Ks) is indicative of positive selection, i.e. advantageous mutations have accumulated during the course of evolution.

The numbers of synonymous and non-synonymous substitutions per synonymous and non-synonymous sites, respectively, have been estimated for the GLI gene family in pairwise comparison by using the Li-Wu-Lu method (Li et al. 1985). Only those codons shared among all species have been considered for the analysis by using the complete deletion option.

GLI3

Within the mammalian lineage, the Ks values for the *GLI3* gene (Table 3.3) range from 0.051 (mouse-rat pairwise comparison) to 0.194 (human-rat). Within fish lineages, the upper level of Ks substitutions approaches saturation level, i.e., Ks >0.4 for zebrafish and tetraodon/*Fugu* comparison. When using pair wise comparisons between members of mammalian and fish lineages, both Ks and Ka values for *GLI3* are in the range of 0.4-0.5.

GLI2

For the *GLI2* gene (Table 3.3), members of mammalian lineages on synonymous sites exhibited a similar evolutionary pattern as *GLI3*, whilst in fish lineages the upper limit of synonymous substitutions at Ks <0.3 did not approach saturation. Mammalian-fish pairwise comparisons indicated a low frequency of synonymous substitutions (0.271-0.368) compared to non-synonymous substitutions.

GLI1

The Ks and corresponding Ka values within mammalian and fish lineages for *GLI1* are lower (Table 3.3), whilst between mammalian and fish lineages the Ks values approached saturation (0.745-

0.858). Corresponding non-synonymous substitution values (0.867-0.947) are higher than for *GLI3* and *GLI2* in pairwise comparison.

Table 3.3. Estimation of Ks and Ka values in pair-wise comparisons

	Human	Mouse	Rat	Tetraodon	Fugu	Zebrafish
GLI3						
Human		0.170 (0.013)	0.185 (0.013)	0.646 (0.027)	0.480 (0.026)	0.487 (0.026)
Mouse	0.161 (0.022)		0.060 (0.007)	0.466 (0.026)	0.491 (0.026)	0.467 (0.025)
Rat	0.194 (0.024)	0.051 (0.011)		0.485 (0.028)	0.507 (0.028)	0.462 (0.025)
Tetraodon	0.426 (0.040)	0.494 (0.046)	0.485 (0.043)		0.106 (0.010)	0.426 (0.024)
Fugu	0.449 (0.043)	0.449 (0.043)	0.507 (0.043)	0.106 (0.015)		0.427 (0.025)
Zebrafish	0.489 (0.043)	0.486 (0.044)	0.462 (0.045)	0.426 (0.040)	0.427 (0.041)	
GLI2						
Human		0.154 (0.013)	0.144 (0.013)	0.477 (0.029)	0.451 (0.029)	0.450 (0.030)
Mouse	0.115 (0.018)		0.068 (0.008)	0.541 (0.030)	0.479 (0.031)	0.462 (0.029)
Rat	0.112 (0.018)	0.032 (0.010)		0.489 (0.029)	0.465 (0.030)	0.470 (0.030)
Tetraodon	0.330 (0.035)	0.365 (0.039)	0.368 (0.039)		0.170 (0.015)	0.422 (0.026)
Fugu	0.271 (0.031)	0.330 (0.036)	0.337 (0.036)	0.111 (0.018)		0.403 (0.025)
Zebrafish	0.306 (0.034)	0.334 (0.037)	0.322 (0.036)	0.296 (0.034)	0.227 (0.028)	
GLI1						
Human		0.160 (0.010)	0.166 (0.011)	0.887 (0.042)	0.871 (0.041)	0.924 (0.046)
Mouse	0.215 (0.021)		0.059 (0.006)	0.906 (0.044)	0.889 (0.042)	0.947 (0.048)
Rat	0.212 (0.021)	0.074 (0.021)		0.901 (0.045)	0.867 (0.041)	0.904 (0.045)
Tetraodon	0.799 (0.059)	0.837 (0.063)	0.828 (0.063)		0.091 (0.007)	0.380 (0.019)
Fugu	0.793 (0.059)	0.868 (0.065)	0.858 (0.068)	0.139 (0.016)		0.366 (0.017)
Zebrafish	0.745 (0.053)	0.788 (0.057)	0.765 (0.056)	0.386 (0.031)	0.377 (0.032)	

Note: The first column and row gives the name of the species for which the pair-wise comparisons were performed. For each member of the GLI gene family (first column), the numbers of synonymous substitutions per synonymous site (Ks) and numbers of non-synonymous substitutions per non-synonymous site (Ka) are presented below and above diagonal, respectively.

3.9.2 Estimation of functional constraints

In order to estimate the selective forces operating on individual *GLI* gene family members following the duplication events, average Ka and Ks values have been estimated for *GLI1*, *GLI*, and *GLI3* genes, both within and between mammalian and fish lineages (Table 3.4). The t-value of difference between average Ka and Ks for each gene has then been used to estimate the significance to which they differ within and between mammalian and fish lineages. Results shown in Table 3.4 suggest that, with the exception of the mammals-fish *GLI2* comparison, there was no significant difference between the average Ka and Ks within or between the two. This indicates a strong trend towards neutrality (Ka/Ks ratio of 1) for substitution rates at synonymous and nonsynonymous sites

for the *GLI* gene family. Only the mammalian-fish comparison for *GLI2* suggests positive selection at 5% significance level ($t = 2.43$, $p < 0.05$).

Table 3.4. Average Ka and Ks values between and within mammalian-fish lineages for *GLI* orthologs

	Ka	Ks	t-value of difference	doublesided p-value	Difference between Means (Ka – Ks)
GLI1					
Mammals-Fish	0.621 ± 0.023 (0.361)	0.579 ± 0.032 (0.371)	0.314	0.7556	non-significant
Mammals	0.130 ± 0.007 (0.056)	0.177 ± 0.014 (0.109)	-0.664	0.5532	non-significant
Fish	0.299 ± 0.011 (0.159)	0.348 ± 0.023 (0.139)	-0.402	0.8013	non-significant
GLI2					
Mammals-Fish	0.374 ± 0.021 (0.152)	0.257 ± 0.021 (0.108)	2.43	0.0206	significant
Mammals	0.133 ± 0.009 (0.054)	0.129 ± 0.015 (0.066)	0.081	0.9586	non-significant
Fish	0.339 ± 0.021 (0.176)	0.341 ± 0.020 (0.178)	-0.014	0.9931	non-significant
GLI3					
Mammals-Fish	0.379 ± 0.025 (0157)	0.372 ± 0.015 (0.162)	0.12	0.9052	non-significant
Mammals	0.155 ± 0.025 (0.070)	0.161 ± 0.025 (0.080)	-0.098	0.9507	non-significant
Fish	0.260 ± 0.011 (0.146)	0.421 ± 0.027 (0.261)	-0.935	0.4285	non-significant

Note: t and p values of pairwise t-tests are also indicated. ± Sign represents standard errors. Standard deviations are enclosed within brackets.

3.9.3 Evolutionary distance between paralogs

To determine the evolutionary rates with which the duplicated genes evolved in each animal tested (human, mouse, rat, *Fugu,* tetraodon, zebrafish), the Tajima relative rate test (Tajima 1993) has been carried out (Table 3.5) on amino acid substitutions on pairs of GLI paralogs, by using the orthologous sequence Ci from *Drosophila* as an outgroup. The Tajima relative rate test determines whether one duplicate has diverged to a greater extent than the other by comparing the sequences of each of the paralogs with that of the ortholog used as the outgroup. The results of this analysis (Table 3.5) indicate that in most cases (14/18) pairs of the GLI paralogs evolve at similar rate in each animal, except in human and mouse, where GLI1 evolved significantly faster ($p < 0.05$) than the GLI2 and

GLI3 counterparts. These findings in relative rate test are in agreement with the notion that paralogs typically evolve at similar rates, without significant asymmetry (Hughes and Hughes 1993) (Kondrashov et al. 2002).

Table 3.5. Tajima relative rate test for the comparison of evolutionary distance between GLI paralogs in different species using the *Drosophila* Ci as an outgroup

Evolutionary Distance	x^2	df	p
Human			
GLI1 vs GLI2	6.43	1	0.011*
GLI1 vs GLI3	5.24	1	0.022*
GLI2 vs GLI3	0.20	1	0.652
Mouse			
GLI1 vs GLI2	9.19	1	0.002*
GLI1 vs GLI3	7.01	1	0.008*
GLI2 vs GLI3	0.17	1	0.676
Rat			
GLI1 vs GLI2	3.33	1	0.068
GLI1 vs GLI3	3.57	1	0.059
GLI2 vs GLI3	0.54	1	0.463
Tetraodon			
GLI1 vs GLI2	3.42	1	0.064
GLI1 vs GLI3	2.14	1	0.143
GLI2 vs GLI3	0.31	1	0.579
Zebrafish			
GLI1 vs GLI2	1.03	1	0.310
GLI1 vs GLI3	0.17	1	0.680
GLI2 vs GLI3	0.01	1	0.920
Fugu			
GLI1 vs GLI2	1.27	1	0.259
GLI1 vs GLI3	0.32	1	0.574
GLI2 vs GLI3	2.04	1	0.153

Note; P-value with "*" symbol represents the situation where GLI1 evolves significantly faster ($p < 0.05$) than the counterpart.

3.10 An insight into the phylogenetic history of *HOX* linked gene families in vertebrates

3.10.1 Phylogenetic analysis

To perform rigorous testing of the 2R hypothesis, which advocates that four-fold paralogy regions in the human HOX-bearing chromosomes might be remnants of polyploidy, phylogenetic analysis of gene families with representatives linked to three or four of the human HOX clusters was conducted (Table 2.4, Fig. 1.20). Gene families with paralogs linked to only two HOX clusters have been left out, because their occurrence is consistent with several alternative explanatory scenarios.

3.10.1.1 Fibrillar Collagen Family – COL

The phylogenetic tree of collagen genes was previously constructed by Bailey and coworkers (Bailey et al. 1997). Their analysis was based on sequence data from very few species (human, mouse and chicken). In this phylogeny, collagen genes on human chromosomes 7, 12 and 17 formed unresolved trichotomy, while genes on chromosome 2 formed an outgroup.

Here, the phylogenetic history of collagen genes was re-analysed by including the sequences from representative members of teleost and tetrapod lineages, thus deriving a clearer picture of evolutionary relationship among members of this family (Fig. 3.34). The phylogenetic tree suggests that duplication events giving rise to members of the vertebrate collagen gene family occurred prior to the actinopterygii-sarcopterygii and after the echinoderms-chordates split. For the COL3A1 gene, the respective time points have not been defined with confidence, because orthologous sequences from actinopterygii are unavailable. Phylogeny indicates with bootstrap support of 83% that COL5A2 was the first molecule of this family to diverge. The remaining family members showed the topology of the form (A)(BCD) (Hughes 1999), i.e. (Hsa2)(Hsa12 Hsa17 Hsa7) with COL3A1 falling outside the cluster of COL2A1, COL1A1, and COL1A2 genes. The branch supporting this pattern received the bootstrap support of 92%.

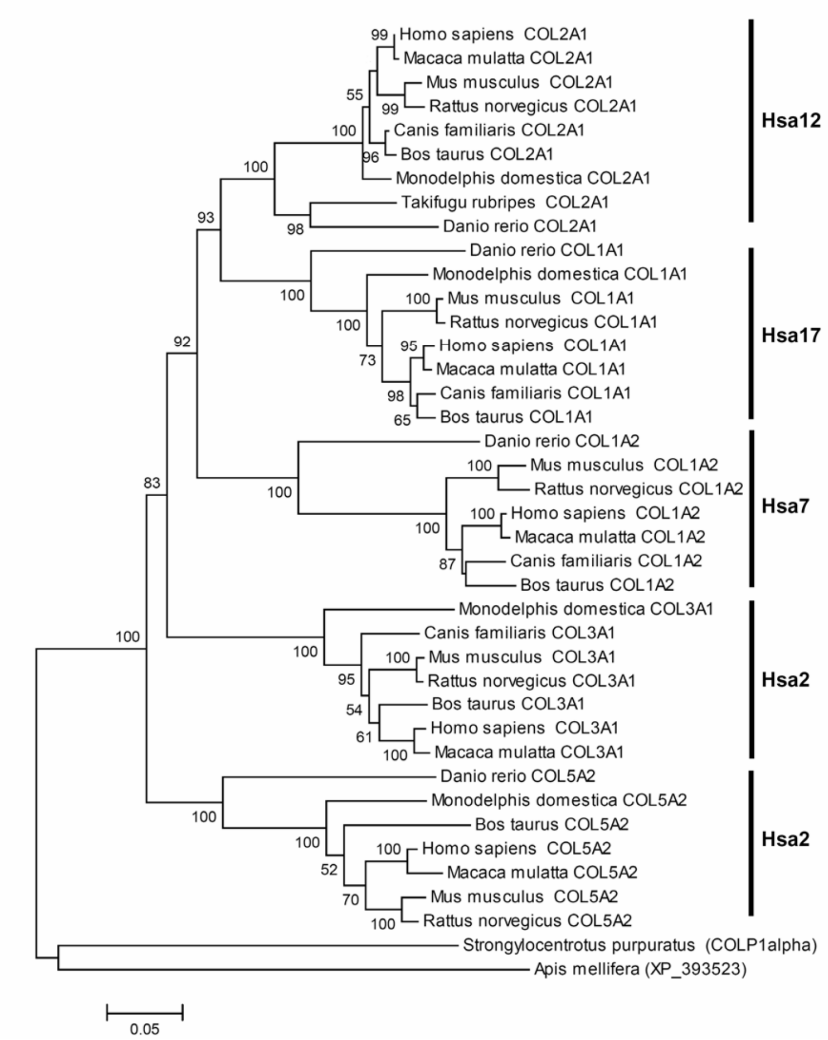


Figure. 3.34. Neighbor-joining tree of the COL family members

Uncorrected *p*-distance was used. Complete-deletion option was used. Numbers on branches represent bootstrap values (based on 1000 replications) supporting that branch; only the values ≥ 50% are presented here. Scale bar shows amino acid substitution per site.

3.10.1.2 ERBB Receptor Protein Tyrosine Kinase – ERBB

For the ERBB family, a topology of the form (A)(BCD), i.e. (Hsa12)(Hsa17 Hsa7 Hsa2), received a strong bootstrap support (97%) with ERBB3 falling outside the cluster of ERBB2, EGFR and ERBB4 (Fig. 3.35).

The phylogenetic tree showed strong evidence of duplications within the time window of deuterostomes-protostomes and actinopterygii-sarcopterygii split.

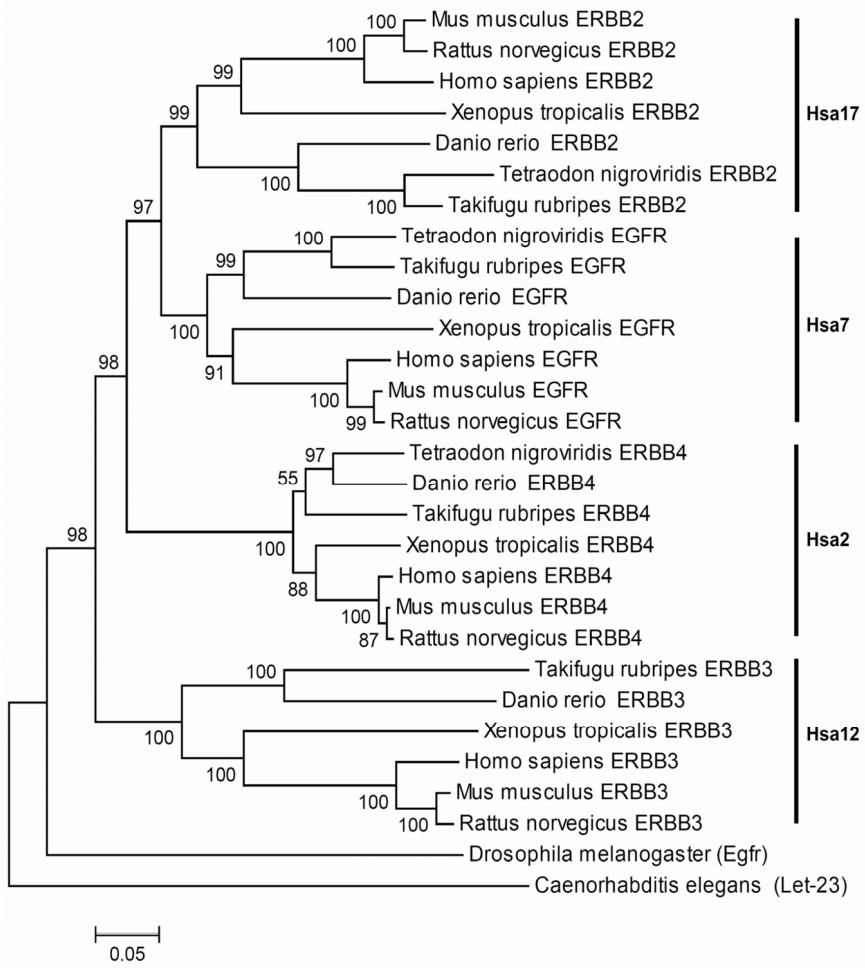


Figure. 3.35. Neighbor-joining tree of the ERBB family
Symbols and parameters are the same as described in Fig. 3.34.

3.10.1.3 Insuline-like Growth Factor Binding Protein – IGFBP

The phylogenetic tree of IGFBP family contained two clusters: (I) a cluster in which the IGFBP5-IGFBP3 genes grouped with IGFBP6, (II) a cluster of vertebrate IGFBP4-IGFBP1 genes grouped with IGFBP2 (Fig. 3.36). The bootstrap support for this pattern was significant, i.e. 99% and 91 % for the two relevant branches. The topology of the vertebrate IGFBP family members is unique in a sense that it can be explained by three, rather than two rounds of gene duplication events early in vertebrate history attributed to (AB)(CD) type gene topology (Hughes 1999). The most parsimonious explanation for this type of topology is: two rounds of whole genome duplication (2R) followed by two independent gene duplication events or three rounds of whole genome duplication followed by two independent gene loss

events. I call this topology an extended form of (AB)(CD) type gene topology in which six genes form two clusters, i.e. (ABC)(DEF).

Phylogeny of vertebrate IGFBP proteins suggests that the gene duplication events giving rise to members of this family have occurred after the urchordates-vertebrates and prior to actinopterygii- sarcopterygii split.

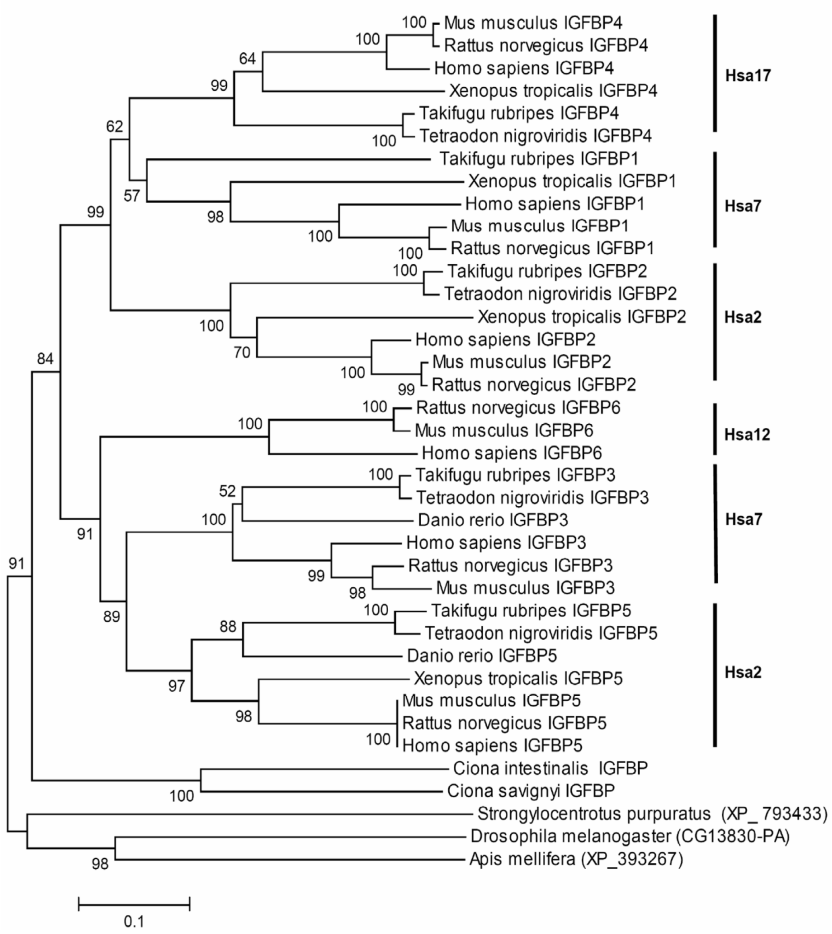


Figure. 3.36. Neighbor-joining tree of the IGFBP family
Symbols and parameters are the same as described in Fig. 3.34.

3.10.1.4 Integrin β – ITGB

In the phylogenetic tree of the integrin β family (Fig. 3.37), vertebrate ITGB5, ITGB3, ITGB6 and ITGB7 genes clustered with homologs from *Drosophila* and sea urchin, indicating that these four members of the integrin β family diverged at least after the divergence of echinoderms and chordates. ITGB4 and ITGB8 genes fell outside the ITGB3-5-6-7 cluster, and homologues from *Drosophila*, and sea urchin. This topology suggests that gene

duplication events giving rise to the ancestor of the ITGB3-5-6-7 cluster may have occurred prior to the divergence of deuterostomes and protostomes.

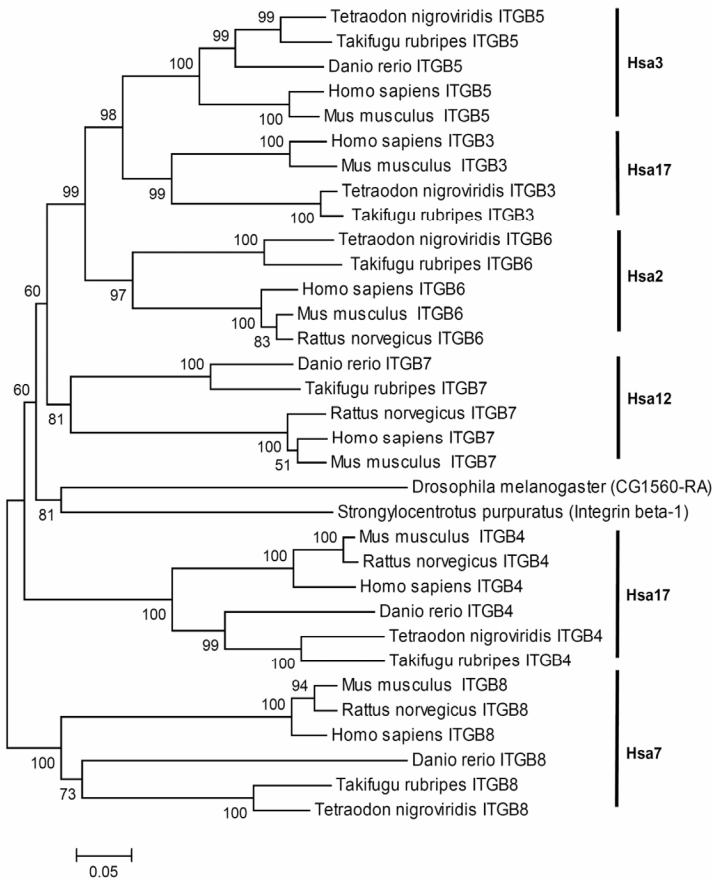


Figure. 3.37. Neighbor-Joining tree of Integrin β chain family .
Symbols and parameters are the same as described in Fig. 3.34.

3.10.1.5 Myosin Light Chain – MYL

The myosin light chain family members formed two major clusters: (I) cluster including vertebrate MYL1, MYL4 and MYL6 genes and homologs from *Drosophila* and *Apis mellifera*, (II) a cluster including vertebrate MYL2, MYL7 and homologs from *Drosophila* and *Apis mellifera* (Fig. 3.38A). Significant bootstrap support, i.e. 100%, for the internal branch separating the two clusters, places the divergence of the ancestors of these two groups prior to the deuterostomes-protostomes split. Subsequent duplications might have occurred early in chordate evolution before the actinopterygii-sarcopterygii divergence.

3.10.1.6 Sp1 C2H2-type Zinc-Finger Protein – SP

In the Sp1 C2H2-type zinc-finger protein family (Fig. 3.38B) a significant internal branch (98% bootstrap support) separated: (I) a cluster containing the vertebrate SP1, SP2, SP3, and SP4 genes showing a topology of the form (A)(BCD), i.e. (Hsa17)(Hsa12 Hsa2 Hsa7) that grouped with a homolog from sea urchin with highly significant bootstrap support, i.e. 98%. (II) The vertebrate SP8 molecule clustered independently with a homolog from sea urchin (95% bootstrap support). The phylogeny suggests that the ancestor of vertebrate SP1-4 and SP8 genes duplicated prior to the divergence of chordates and echinoderms.

3.10.1.7 Zinc-Finger Protein-Subfamily 1A – ZNFN1A

The vertebrate members of ZNFN1A family showed a topology of the form (A)(BCD), i.e. (Hsa12)(Hsa7 Hsa17 Hsa2), with ZNFN1A4 clustered outside the other three vertebrate genes. The branch supporting this pattern received bootstrap support of 90% (Fig. 3.38C). The topology of the phylogenetic tree indicated that the gene duplications giving rise to ZNFN1A family members occurred within the time window of echinoderms-chordates and actinopterygii-sarcopterygii split.

3.10.1.8 Anion Exchanger – SLC4A (AE)

The phylogenetic tree of SLC4A genes (Fig. 3.38D) is divided into two major clusters. Cluster-1 includes vertebrate members SLC4A1, SLC4A2, SLC4A3, and a homolog from *Drosophila*; cluster-2 includes SLC4A5, SLC4A8, SLC4A10, and homologs from *Drosophila* and *C. elegans*. The internal branch separating the two clusters received highly significant (100%) bootstrap support. The topology suggests that gene duplication events giving rise to ancestors of cluster-1 and cluster-2 occurred prior to deuterostomes-protostomes divergence. Phylogeny further indicates that the mammalian SLC4A8 and SLC4A10 genes arose through the duplication of an SLC4A8-like ancestor in the tetrapod lineage at least before the divergence of Euarchontoglires from Laurasiatheria, and the branch supporting this pattern received 91% bootstrap support.

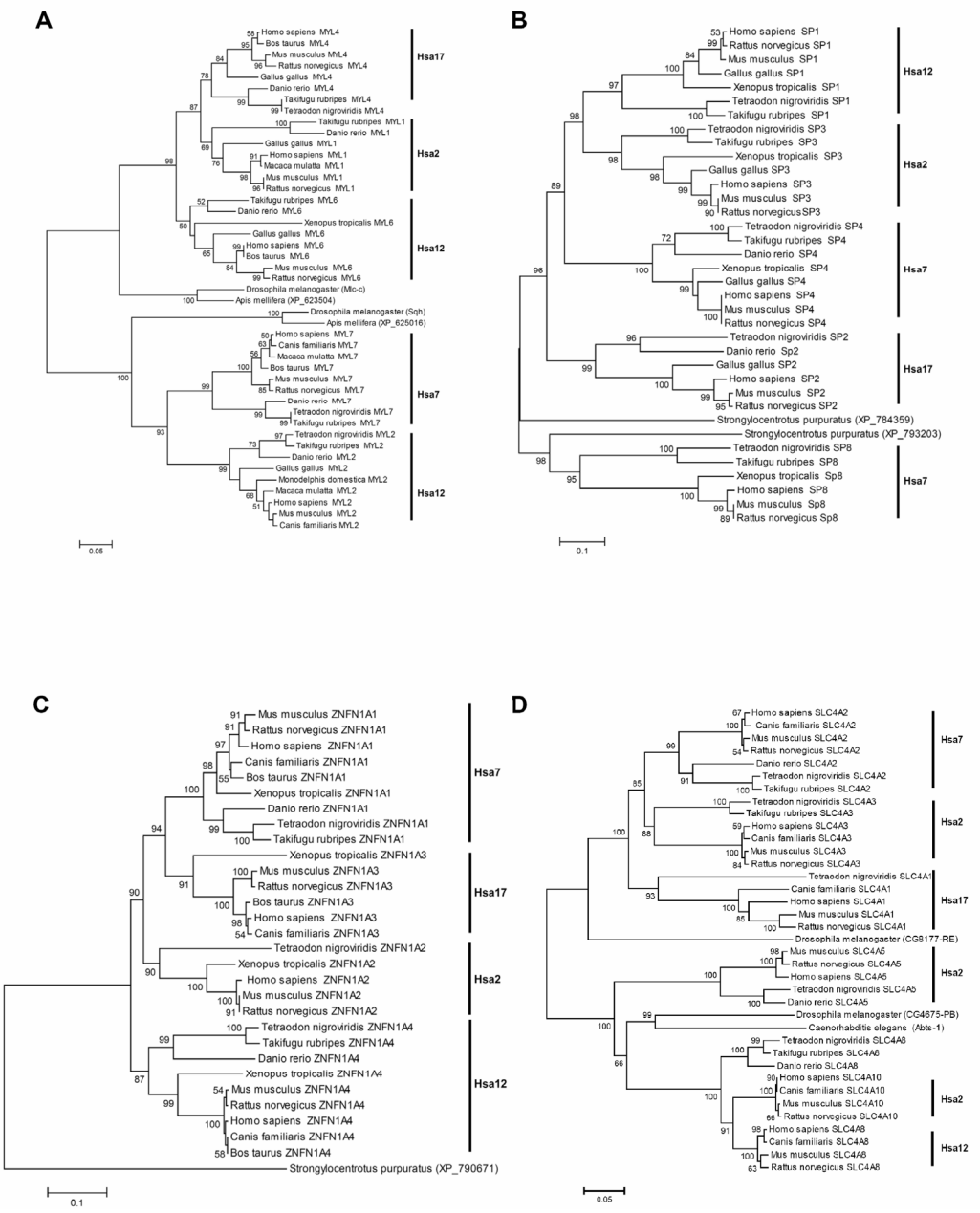


Figure. 3.38. Neighbor-joining tree of the (A) Myosin light chain family, (B) SP family, (C) ZNFN1A family, (D) SLC4A family

Symbols and parameters are the same as described in Fig. 3.34.

3.10.1.9 GLI Zinc-Finger protein – GLI

The phylogenetic tree indicates that the GLI1, GLI2, and GLI3 genes diverged after the separation of urchordates from vertebrates and before the divergence of tetrapods and bony

fishes (Fig. 3.39A). The phylogeny shows a topology of the from (A)(BC), i.e. (Hsa12)(Hsa7 Hsa2) with highly significant (100%) bootstrap support.

3.10.1.10 Hedgehog – HH

Vertebrate HH family members showed a topology of the form (A)(BC), i.e. (Hsa12)(Hsa7 Hsa2), with DHH falling outside the SHH-IHH cluster. The branch supporting this pattern received significant (97%) bootstrap support (Fig. 3.39B). Phylogeny attributed the birth of the vertebrate HH family members to duplications which occurred within the time window of the cephalochordates-vertebrates and tetrapods- fishes split.

3.10.1.11 Inhibin – INHB

The topology of vertebrate inhibin genes (Fig. 3.39C) is similar to the HH and GLI families, i.e. (Hsa12)(Hsa7 Hsa2) with 93% bootstrap support. Furthermore, the phylogenetic tree indicates that inhibin paralogs on Hsa12, i.e. INHBC and INHBE originated by a duplication event in tetrapod lineage after its divergence from bony fishes. The branch supporting this pattern received significant (96%) bootstrap support.

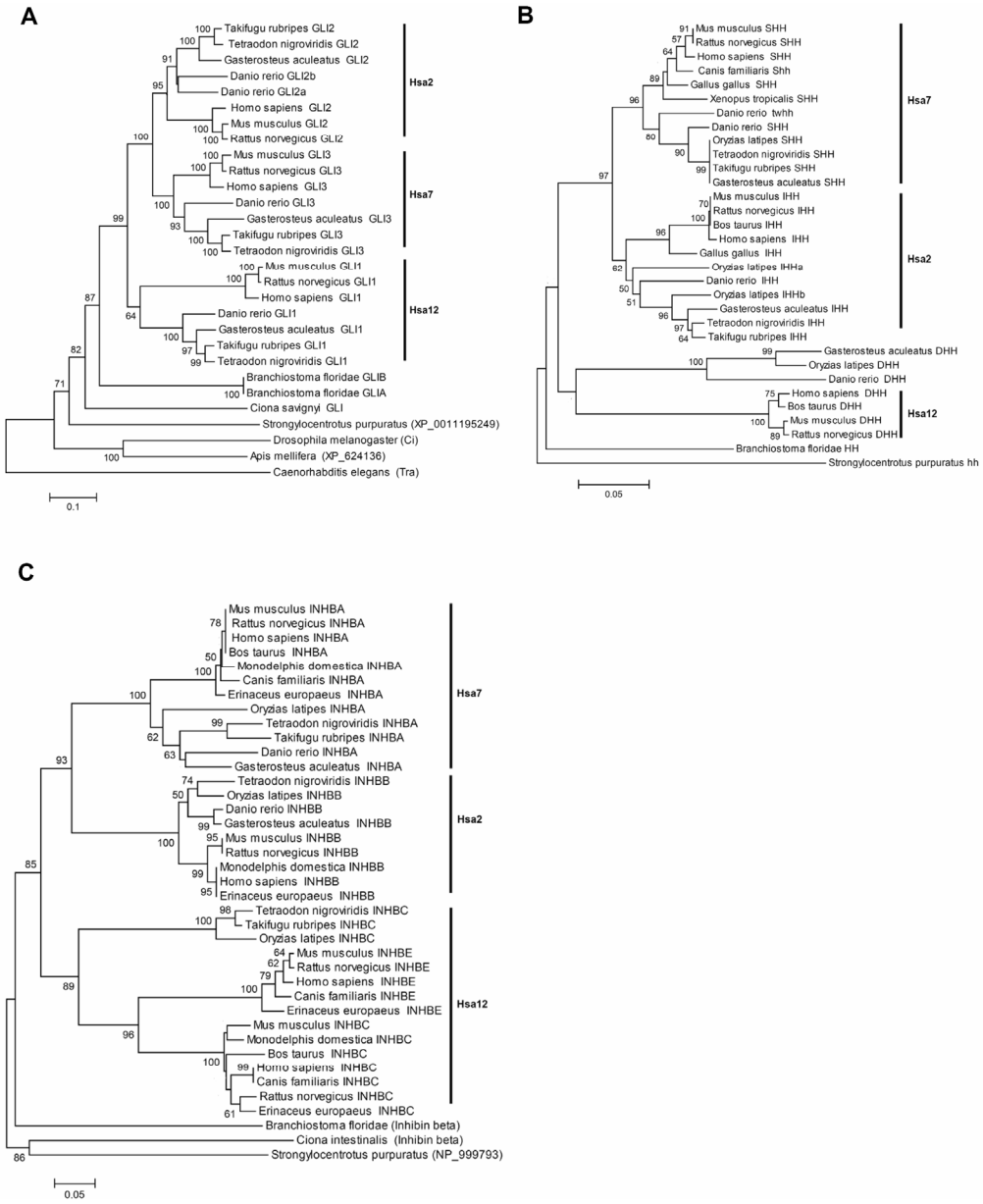


Figure. 3.39. Neighbor-joining tree of the (A) GLI family, (B) Hedgehog family, and (C) Inhibin family.

Symbols and parameters are the same as described in Fig. 3.34.

3.10.2 Estimation of co-duplication events

Given the phylogenetic data, I next sought to determine which genes could have duplicated simultaneously. To test this, the the topology comparison approach was adopted (Hughes et al. 2001). Genes were selected from those portions of each phylogeny, where there was a strong statistical support for duplication events within the time window of vertebrates-

invertebrates and tetrapods-fishes split (Fig. 3.40, Table 3.6). Furthermore the published phylogeny of vertebrate HOX clusters (Zhang and Nei 1996) was also included in this test.

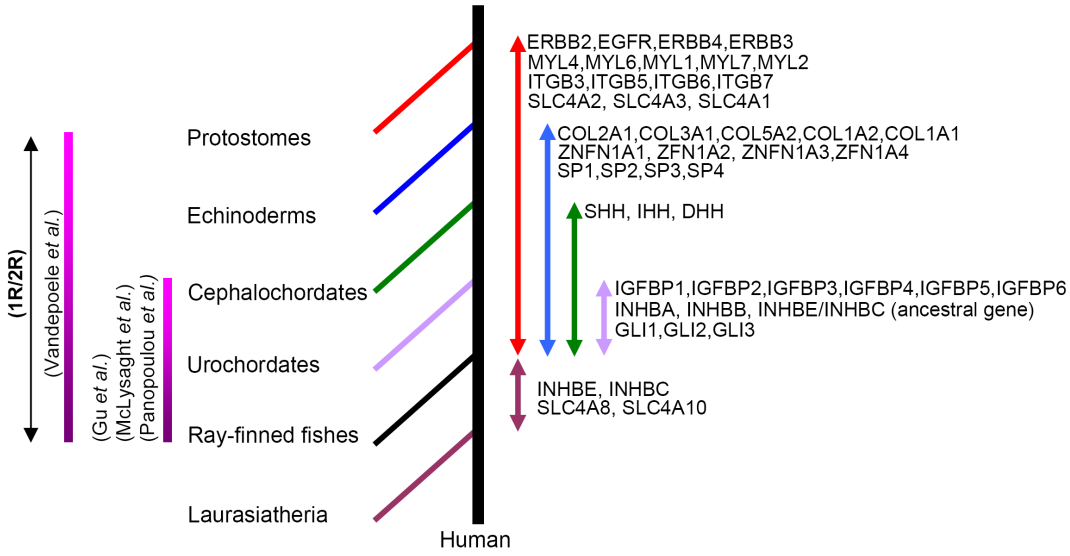


Figure. 3.40. Members of HOX linked gene families that have arisen early in vertebrate history

Order of branching within phylogenetic trees was used to estimate the time windows (double headed arrows on the right) of gene duplication events relative to major cladogenetic events. For each family the lower limit of time window was defined from fish-tetrapod split and the upper limit from the branching order of available closest invertebrate ancestral sequence (Protostomes: *Drosophila*, *Apis mellifera*; Echinoderm: Sea Urchin; Cephalochordates: Amphioxus; Urchordate: *Ciona intestinalis*, *Ciona savignyi*). The INHBE, INHBC and SLC4A8, SLC4A10 genes arose after the fish-tetrapods split. Previously proposed timing (Gu et al. 2002; McLysaght et al. 2002; Panopoulou et al. 2003; Vandepoele et al. 2004) of extensive gene duplications during early chordate evolution is indicated on the left of the diagram.

The topology of the type where genes on chromosomes 7 and 2 clustered together and the gene on chromosome 12 formed an outgroup (Table 3.6) depicts the simultaneous duplication of members of five gene families, i.e. GLI, HH, INHB, IGFBP (cluster-1), and SLC4A. The third member of the SLC4A family, i.e. SLC4A1, that forms an outgroup to the SLC4A2-SLC4A3 cluster, is on a different chromosome (Hsa17), suggesting that an independent translocation event followed the co-duplication.

The topology of the type where genes on Hsa7 and Hsa17 clustered together, while the gene on Hsa2 branched next, and the gene on Hsa12 formed an outgroup (Table 3.6), is suggestive of another gene-cluster duplication event involving the members of ERBB, ZNFN1A, and IGFBP (cluster-2) gene families. In addition, the genes showing the topology of the type (Hsa12) (Hsa7 Hsa17 Hsa2) maintained exactly the same order on the respective

chromosomal segments, with ZNFN1A genes flanked by ERBB and IGFBP family members (Fig. 1.20). This reflects a conservation of gene order following co-duplications.

In the previously published phylogeny of vertebrate HOX clusters, HOXC and HOXD are grouped together, while the branching order of HOXA and HOXB is unresolved; two alternative topologies (((HOXC HOXD)HOXA)HOXB) and ((HOXC HOXD)(HOXA HOXB)) are equally probable (Zhang and Nei 1996). Within the phylogeny of the SP family, the branching order of SP1, SP2, SP3, and SP4 genes is congruent with one of the two proposed alternative phylogenies of HOX clusters (Table 3.6). Consistent with the compatibility in their tree topologies, each of the relevant SP genes is closely linked with the HOX cluster (Fig. 1.20), with human SP1 gene mapping at approximately 526 kb centromeric to HOXC, SP2; at ~ 614 kb centromeric to HOXB, SP3; at ~2 Mb centromeric to HOXD, and SP4; at ~5 Mb telomeric to HOXA. This implies that HOX linked SP genes share the similar evolutionary history as the HOX clusters and have arisen through the same duplication events that led to the HOX clusters.

The phylogenies of the integrin beta chain and myosine light chain families, where the vertebrate genes on Hsa17 and Hsa2 clustered together and the gene on Hsa12 formed an outgroup (Table 3.6) revealed a fourth simultaneous duplication event. The fact that ITGB3 on Hsa17 grouped with ITGB5 on Hsa3 suggests that an independent translocation event followed the duplication of their ancestor after its divergence from the ITGB6 gene (Fig. 3.37).

The phylogeny of collagen genes showed a different topology (Table 3.6) which is inconsistent with their having duplicated concomitantly with members of any other gene family that is included in the current study.

Table 3.6: Summary of the Phylogenetic Analysis of Gene Families

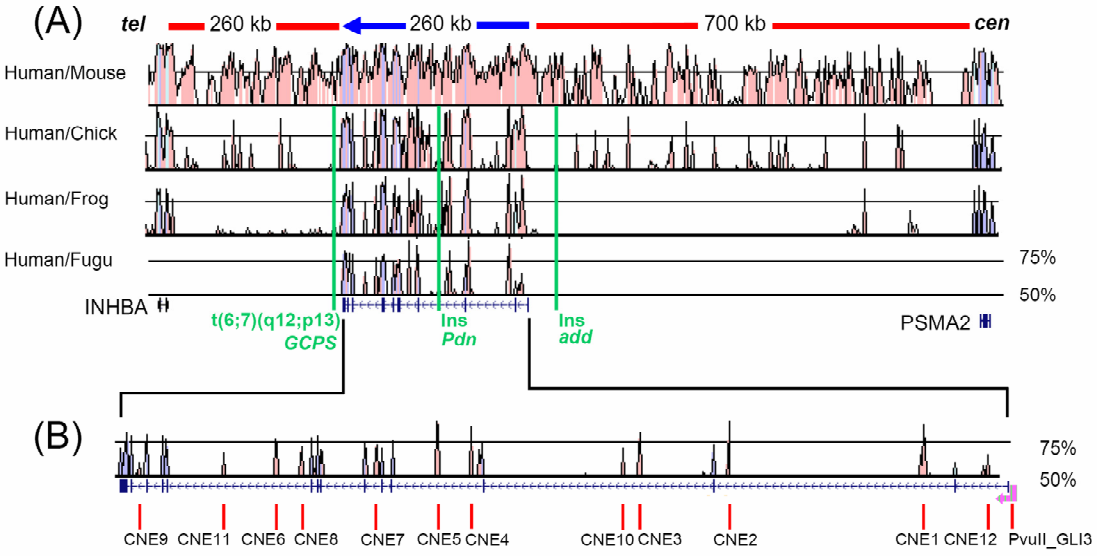
Family Name	Hsa2	Hsa7/3*	Hsa12	Hsa17	Consistency with *HOX* Phylogeny	Topology
ERBB	ERBB4	EGFR	ERBB3	ERBB2	-	(((17, 7) 2) 12) 97,98
Collagen	COL3A1 COL5A2	COL1A2	COL2A1	COL1A1	-	((((12,17)7)2)2) 93,92,83
IGFBP	IGFBP2 IGFBP5	IGFBP1 IGFBP3	IGFBP6	IGFBP4	-	((17, 7)2) ((7, 2)12) 99,91
INTB	ITGB6	ITGB5*	ITGB7	ITGB3	-	(((3, 17)2)12) 98,99
MYL	MYL1	-	MYL6	MYL4	-	((17, 2)12) 87
SP	Sp3	Sp4	Sp1	Sp2	Yes	(((12, 2) 7) 17) 98,89
ZNFN1A	ZNFN1A2	ZNFN1A1	ZNFN1A4	ZNFN1A3	-	(((7, 17) 2) 12) 94,90
INHB	INHBB	INHBA	INHBC INHBE	-	-	((7, 2)12) 93
SLC4A	SLC4A3	SLC4A2	-	SLC4A1	-	((7, 2)17) 85
GLI	GLI2	GLI3	GLI1	-	-	((7, 2)12) 99
HH	IHH	SHH	DHH		-	((7, 2)12) 97

For each gene family the chromosomal location and topologies (in the Newick format) of those genes are given, which arose through duplications after the invertebrates-vertebrates split and before the tetrapods-fishes divergence. The percentage bootstrap support of the internal branches is given below each relevant topology.

DISCUSSION

4.1 Comparison of the genomic architecture in and around *GLI3* reveals an ancient gene regulatory network (AGRN) within its introns

Human *GLI3* extends over 260 kb on chromosome 7p14.1 (Fig. 4.1A), a gene poor region, and is flanked by ~260 kb and ~700 kb intergenic intervals (Scherer et al. 2003). In order to investigate the genomic architecture in and around *GLI3,* a comparative genomic analysis was performed by multispecies sequence comparison (Fig. 4.1) of a ~ 1220 kb interval encompassing *GLI3*. Orthologous sequences for the analysis were imported from ENSEMBL genome browser (http://www.ensembl.org), and to perform a pairwise comparison of sequences from several species the computer program Shuffle-LAGAN (Brudno et al. 2003a) was employed, because of its ability to align long genomic sequences, while allowing for rearrangements events such as translocations, inversions, duplications or a combination of above .



Human cell lines	A/R	A/R	A/R	–	A/R	A/R	R	A/R	R	A	A/R	A/R	–
Zebrafish	(+)	(+)	(+)	-	(+)	(—)	(—)	(+)	(—)	(+)	(+)	-	(+)
Mouse	(+)	(+)	(+)	-	(+)	-	-	(+)	-	(+)	(+)	-	(+)
Chick limb bud	(—)	(+)	-	-	-	-	-	(—)	-	-	(—)	-	-

Figure. 4.1. Comparative sequence analysis of the *GLI3* locus detects conserved non-coding sequence elements.

(A) Sequence alignments of the genomic interval encompassing the human *GLI3* locus and the flanking genes *INHBA* and *PSMA2* with orthologous counterparts from representative members of rodent, bird, amphibian, and fish lineages. These are shown as SLAGAN derived VISTA representations. Conserved coding sequences are depicted in blue, and conserved non-coding sequences are in pink. Criteria of alignment were 60 bp window and 50% conservation cutoff. Conservation between human and *Fugu* is restricted to the *GLI3* gene. Red bars above the conservation plot depict the approximate length of intergenic regions flanking human *GLI3*. The blue arrow shows the length of the *GLI3* gene and the direction of transcription. A graphical representation of exons and introns of *GLI3* is shown below the homology plot. Green vertical lines indicate the positions of alterations affecting the genomic structure of the locus which result in loss of *GLI3* function: a translocation event associated with Greig cephalopolysyndactyly syndrome (GCPS) (Radhakrishna et al. 1999), and two insertions (ins) in mouse mutants anterior digit pattern deformity (*add)* (Pohl et al. 1990; van der Hoeven et al. 1993) and polydactyly Nagoya (*Pdn)* (Ueta et al. 2004).
(B) Magnified view of the human/*Fugu* conservation plot and the genomic structure of *GLI3*. The red vertical bars below the plot show the position of human/*Fugu* highly conserved non-coding sequence elements (CNEs) that were functionally tested as putative enhancers. The promoter proximal element of *GLI3* (GLI3_PvuII) that has been tested functionally is shown with pink arrowhead; direction of arrow indicates the direction of transcription. Underneath is shown in tabular form the *in vitro* (human cell lines) or *in vivo* (model organisms; mouse, zebrafish and chick) activity of these elements. Dual nature and exclusively repressory elements are represented by "A/R" (activator/repressor) and "R" symbol, respectively. The (+) sign indicates the elements which induced reporter expression in mouse, zebrafish or chick embryos while a (–) sign indicates those which could not drive reporter expression in these model organisms.

This comparative sequence analysis has revealed extensive non-coding tetrapod/teleost sequence conservation, distributed throughout *GLI3* intronic intervals, but in contrast to the findings in which long intergenic intervals have been associated with a tetrapod/teleost specific ancient gene regulatory network (AGRN) (Nobrega et al. 2003), tetrapod/teleost synteny at the *GLI3* locus breaks immediately before and after the fist and last codon, respectively. However, close inspection of mouse, chick, and frog orthologous intervals revealed conserved synteny across flanking 5′ and 3′ intergenic intervals of *GLI3*, as well as the centromeric and telomeric gene neighborhood (Fig. 4.2).

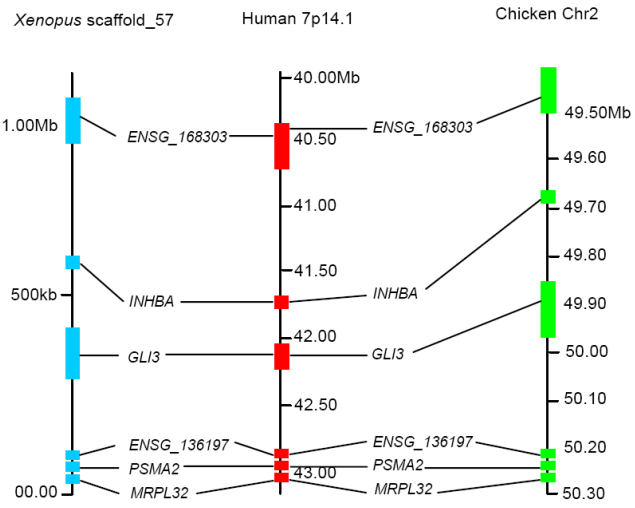


Figure. 4.2. Conserved synteny and gene order among human 7p14.1, chicken Chr 2, and Xenopus scaffold_57

The gene content of the human interval is depicted by HUGO gene symbols or ENSEMBL gene ID. Scale is in megabase for human and chicken regions while a kilobase scale has been used to show the Xenopus part of scaffold _57.

As a consequence of the fact that conservation of non-coding tetrapod/teleost sequences is strictly bracketed within the first and the last exon of *GLI3,* the AGRN is kept within its introns, rendering *GLI3* out of proximal and distal genomic context. The flanking intervals evolved independently in the tetrapod lineage after its divergence from teleosts (450 Mya) having no shared functional constraints between two lineages. However, loss of tetrapod/teleost sequence conservation may not rule out the existence of gene regulatory elements flanking *GLI3*, because the conservation of orthologous sequences flanking *GLI3* 5´ and 3´ within human, chick and frog is indicative of strong selective constraints which are shared by the tetrapod lineage and which are operating on the region after its divergence from teleost fish. This conservation might be attributed to distant regulatory elements, controlling the expression of *GLI3* or of flanking genes on either side.

Evidence in favor of widespread distribution of *GLI3* specific regulatory elements comes from reports of rearrangement events (Fig. 4.1) resulting in phenotypes similar to those reported for a direct consequence of errors within the coding interval of *GLI3.* A translocation, t(6;7)(q27;p13), truncates chromosome 7p14 about 10 kb downstream of the last exon results in a GCPS phenotype characteristic for functional deficiency of one GLI3 allele (Kruger et al. 1989) (Fig. 4.1). Silencing of the intact *GLI3* gene in this case could be caused by loss of cis-regulatory sequences distal to the breakpoint. In the mouse, a transgene insertion ~ 64 kb upstream of *Gli3* is associated with the phenotype of anterior digit deformity (add) (Pohl et al. 1990; van der Hoeven et al. 1993) (Fig. 4.1). Here, the function of murine Gli3-enhancers located upstream of the insertion site may be disturbed. Thus, it is possible that *GLI3* might be among the genes regulated by distant enhancers. A further *Gli3* allele, the mouse mutant Pdn, results from insertion of a transposon into intron 3 (Fig. 4.1) (Ueta et al. 2004). In those mice Gli3 expression appears to be possible, though at a reduced level. Based on these clinical observations, *GLI3* regulatory elements might be found up- or downstream of the gene or within the introns.

Algorithms for the prediction of enhancers determining the temporal and spatial expression of human genes are increasingly powerful (Bejerano et al. 2005), however, sound predictions of *GLI3* regulatory signatures have not yet been reported.

The region around *GLI3* is prohibitively large for using the painstaking strategy of stepwise deletions in reporter gene assays. Recently, it has been reported that there is a considerable overlap between experimentally verified enhancer elements and non-coding sequence elements (CNEs), that are evolutionarily conserved between distantly related species such as humans and the pufferfish (Pennacchio et al. 2006). This suggests that CNEs around or within a gene are promising candidates for enhancers of expression. Different levels

of stringency have been applied for the definition of CNEs (Bejerano et al. 2004; Sanges et al. 2006), mostly with the intention to select a manageable number of candidate elements rather than with a biologically based rationale.

By employing multispecies sequence alignment we identified an ancient (tetrapod-teleost conserved) non-coding architecture within the introns of *GLI3*. The ancient, human/fish conserved signatures are embedded in larger sequence domains conserved in evolutionarily more recent species such as frog, chick or mouse (Fig. 4.1). To test for possible enhancers of expression, 11 human/mouse CNEs were chosen encompassing >60 bp DNA tracks with more than 50% sequence similarity between human and *Fugu*. These candidate elements represent sequence that is under ancient, strong evolutionary constraint operating to maintain a DNA sequence signature.

4.2 *In-vitro* regulatory activity of intra-*GLI3* CNEs is cell type specific

The majority of intra-*GLI3* CNEs (8/11) exhibited a cellular context dependent dual nature. In the endogenous *GLI3* expressing environment (H661) they functioned as activators whilst in the GLI3 negative (H441) cellular context they actively repressed the transcription (Fig. 3.13). This differential activity is strong evidence in favor of assigning a *GLI3*-specific regulatory potential to these CNEs. Similar context dependent dual-nature regulatory activity is known for other transcription factors (Grice et al. 2005). The *in vitro* investigation also revealed two CNEs that had a repressing potential, even in a GLI3 positive cellular context (Fig. 3.13). The ultraconserved element CNE2 functioned as an activator only in a GLI3 positive context and it was unable to influence the reporter activity in the endogenous GLI3 negative environment (Paparidis et al. 2007).

The most plausible scenario to explain the dual nature of a sub-set of intra-*GLI3* enhancers could be the interaction of each CNE with different subsets of trans-acting factors (either activators or repressors of transcription) in a cellular context dependent manner (Hersh and Carroll 2005), whilst elements with repressing potential, even in a GLI3 positive context, suggest the existence of context independent regulation.

4.3 *In-vitro* deletion analysis defines functional modules within CNE1, 5 and 6

A subset of the CNEs, each associated with unique sequence features, was employed as potential enhancers in transient transfection assays in H661 cells, to see if core elements conserved in human-*Fugu* represent functionally critical regulatory modules. CNE1 spans a human/*Fugu* conserved region of exceptionally extended length, 935bp, and embeds within it

a highly constraint interval of 125bp almost 100% conserved down to chick, while depicting a 92% conservation in human/fish comparison (Fig. 3.14). CNE5 showed 85% identity over 255bp in human/fish sequence comparison and encompasses human to fish 100% conserved contiguous binding sites for developmentally important TFs PBX1, PAX2 and MEIS1 (Fig. 3.15). CNE6 contains a small, moderately conserved human/fish track of 179bp, within a human/mouse 862bp track with overall 87% conservation (Fig. 3.16).

Considering the known degeneracy of transcription factor binding target sites (Stormo 2000), the high conservation of the 125bp track within CNE1 over the course of such an extreme phylogenetic separation (human/fish) is unexpected (Boffelli et al. 2004). A previous *in vivo* attempt to resolve the significance of a tetrapod-conserved non-coding sequence element encompassing a human-*Fugu* conserved region around the human *DACH* gene had revealed that alone, neither a human-*Fugu* conserved core nor the less conserved flanking region functioned as activators (Poulin et al. 2005). The authors concluded that either the assay was not sufficiently sensitive or the core element might have an unknown biological function. In contrast, our *in vitro* deletion analysis with CNE1 provides evidence in favour of a quantitative participation of the conserved core to the overall activity of the enhancer, and suggests that this module embedded within CNE1 is essentially gene regulatory in function (Fig. 3.14). The fact that the 125 bp sequence was unable to show any detectable activity in isolation reflects that this module is necessary but not sufficient to uphold enhancer function.

Similarly, the manipulation of CNE5 revealed that the deleted highly conserved element encompassing predicted contiguous TFBSs for developmental regulators *PBX1*, *PAX2* and *MEIS1* is necessary but not sufficient for the activating potential of this site (Fig. 3.15).

The *in vitro* reporter gene expression data from deletion constructs of CNE6 suggest the existence of tetrapod specific functional constraints in the vicinity of an ancient fish specific element (Fig. 3.16). It appears that the overall enhancer activity is determined by the combinatorial affect of the ancient and the more recent sequences (Fig. 3.16).

In all 3 examples, the excised elements had no activating potential when analyzed without the flanking sequences. However, we conclude from the sizeable reduction in activating potential in the absence of the core, that the human-*Fugu* conserved modules within the *GLI3*-CNEs are critically involved in transcriptional regulation. Mutagenesis of the predicted binding sites could show if transcription factors are involved in this function. The fact that flanking intervals of the human-*Fugu* conserved sequence elements contribute to the activity of the elements suggests, that after the divergence of tetrapod-teleost lineages (450 Million years ago) there was a progressive gain of novel functions centered around an ancient enhancer element (Fig. 4.3). This provided an opportunity for fine-tuning of gene expression

differentially in the tetrapod lineage, congruent with their complex developmental and anatomical needs.

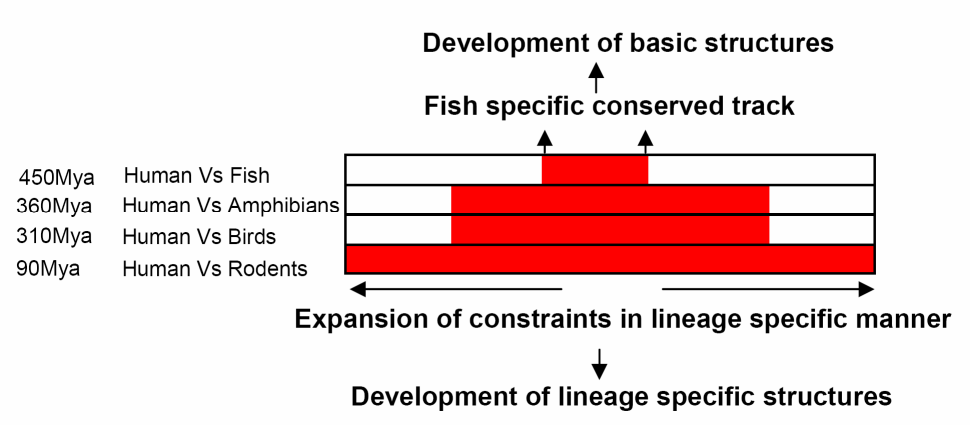


Figure. 4.3. Progressive gain of novel regulatory components around an ancient enhancer element

The red colored boxes show sequence conservation of an enhancer element among relevant species (pairs shown at the left). The human sequences around human-fish moderately conserved enhancer elements (red block at the top) progressively depict a more expanded span of negative selection while the divergence time (given at left in Million years) between the human and non-human bony vertebrates (in this case fish, amphibian, bird and rodent lineages are shown) used in sequence comparison gets shorter. This pattern suggests that on the basis of ancient vertebrate *cis*-acting regulatory units the complexity was increased gradually in tetrapods to fine tune gene expression differentially in distinct lineages, congruent with their specific developmental and anatomical needs in terrestrial ecosystems.

4.4 Can transcription factor binding sites within CNEs explain their evolutionary conservation?

A possible restraint causing the maintenance of CNEs involved in gene regulation throughout vertebrates could be a strict combinatorial code of TFBSs where order and distance are critical. CNEs were screened for human-*Fugu* conserved putative TFBSs using the computer programs Consite and rVista v 2. In order to increase the sensitivity and to reduce the number of false positives, the TFBSs motif searches were combined with phylogenetic footprinting of CNEs across distantly related species (Loots and Ovcharenko 2004; Sandelin et al. 2004). In each of the 11 sequences, human-*Fugu* conserved binding sites were identified for a number of developmental regulators. The prediction of binding sites for established developmental regulators under the highly stringent criteria in each of the tetrapod-teleost conserved intra-*GLI3* sequence tracks corroborates the conclusion from our experiments that the ancient elements contribute to the activity as an enhancer. However, TFBSs are known to allow considerable degeneracy, and their overall density across each individual CNE is low. Unless strict maintenance of a combination of specific TFBSs and

flanking sequence is required to retain tissue specifity of enhancer action, these sites may not contribute to the major constraint responsible for conservation of non-coding elements throughout evolution.

4.5 Intra-*GLI3* CNEs show tissue specific regulatory activity in zebrafish embryos

In order to address the *in vivo* role of *GLI3*-associated conserved non-coding elements, a medium throughput strategy was selected, employing transient reporter gene expression from the human *β*-globin promoter under the influence of a putative enhancer element in zebrafish embryos (Woolfe et al. 2005). This approach exploiting the transparency and rapid development of zebrafish embryos has recently shown its immense potential for functionally testing enhancer elements among conserved non-coding regions (Goode et al. 2005; McEwen et al. 2006; Muller et al. 1999; Woolfe et al. 2005). Our results indicate that the regulatory potential of most of the human CNEs defined in transient transfection assays of human cell cultures is similarly present in fish embryos. There is also a correlation between both enhancer and repressor activity *in vitro* and *in vivo*. Thus, we present evidence that both the sequence and the regulatory characteristics of *cis*-acting elements are conserved throughout evolution, from teleosts to man.

In mouse, GLI3 plays a prominent role in development of brain, ear, eye, craniofacial structures, limb and lung, and is also expressed in heart, kidney, skeletal muscles, fetal blood cells, epidermal cell layer of skin and other tissues (Mouse Genome Informatics http://www.informatics.jax.org). Zebrafish *gli3* is reported to be expressed in brain, dorsal spinal cord neurons, eye, and pectoral fin bud (Zebrafish Information Network; http://zfin.org) (Tyurina et al. 2005; Vanderlaan et al. 2005). However, exhaustive expression patterns throughout different stages of development have not been published.

A number of the positions in which transgene expression is observed coincide with known sites of GLI3 activity. For example CNE1 and CNE2 drives GFP expression predominantly in various subdivisions of the CNS, CNE10 activity was most frequent in the eye, pericardial region, lower jaw primordia and skin cells, CNE6 activity was more specific to hindbrain/spinal cord boundary neurons, muscle fibers and blood cell, and CNE11-driven reporter expression was largely restricted to cardiac chambers, skin cells and muscle fibers. Interestingly, CNE11 also induced GFP expression with low frequency within pectoral fins at day 3 of development, which is consistent with the reported timing of zebrafish *gli3* expression in this tissue (Tyurina et al. 2005). The lack of specificity in the enhancer activity of CNE7 can be attributed to the absence of normal genomic context (Gomez-Skarmeta et al. 2006; Woolfe et al. 2005) which might impose constraints to dictate the precise endogenous

activity of an enhancer, while in our analysis we are testing the inducing capability of putative enhancers out of endogenous genomic context. It can be seen that functional redundancy with respect to the site of expression was evident for all regulatory elements, a notion concordant with findings in other genes (Fisher et al. 2006). Some cell populations such as heart, the pericardial region, blood cells, muscle fibers, skin, and lower jaw primordia are domains of *Gli3* expression in mouse but not so far described in zebrafish. However, GLI3 functions appear to be conserved in mouse and zebrafish (Tyurina et al. 2005). Therefore, the expression of g*li3* in zebrafish might be more extensive than reported so far. We observed expression in domains of the embryo where *gli3* is expressed neither in zebrafish nor in mouse. For example, CNE9 directed expression predominantly to the notochord, which is inconsistent with the reported endogenous *GLI3* expression in either species. The unexpected finding of a CNE within *GLI3*, which directs reporter gene expression at a site where GLI3 itself is never observed, stresses the importance of normal genomic context for the function of regulatory elements, as had been concluded by previous studies (Fisher et al. 2006; Gomez-Skarmeta et al. 2006; Woolfe et al. 2005).

4.6 Only CNE 11 could evoke reporter gene expression in chick limb-bud

The *GLI3* gene product is critical for growth and patterning of limb along anterior (thumb) – posterior (little finger) and proximal – distal axis. In order to test whether any of these anciently conserved of intra-*GLI3* CNEs regulates the expression of this gene in the limb bud of tetrapods, we selected four of these elements (CNE1, 9, 10, and 11) to analyze their reporter inducing capability in chick limb-bud.

Consistent with zebrafish *in vivo* data, CNE11 shows weak enhancer activity in the chicken limb, bud but we could not detect any enhancer activity with CNEs 1, 9 and 10. Using the same assay it has been demonstrated previously that conserved non-coding elements downstream of the homeobox gene *SHOX* have enhancer activity (Sabherwal et al. 2007). In this case, three out of the eight CNEs tested showed enhancer activity indicating that some but not all conserved non-coding elements act as enhancers. The difference between this experiment and the previous experiments with SHOX is that, at later stages of development *Gli3* has a more restricted expression pattern and lower level of expression than *SHOX*. This may reduce the chance of introducing a putative enhancer construct into a *Gli3* expressing region of the limb bud and therefore make it more difficult to detect enhancer activity. However, introduction of CNE11 into the limb at an earlier stage when *Gli3* expression is more widespread did not show any enhancer activity.

Taken together, the data suggest that, even though the CNE11 function was in concurrence with *Gli3* expression within nascent limb bud of chick, it recapitulated only a part of the entire repertoire of endogenous *Gli3* expression there, i.e. its activity was late in time and was restricted to a proximal portion of the limb-bud. This data signals, that there might be enhancer regions other than CNE1, 9 and 10, which function in conjunction with CNE11 to trigger the full spectrum of spatiotemporal aspects of normal *Gli3* expression in limb-bud.

4.7 The evolutionary conserved human cis-regulators involved in the mediation of spatiotemporally distinct sub-domains of Gli3 expression in mouse

After confirming the regulatory potential of human *GLI3* associated, highly conserved intronic sequences in human cell lines, zebrafish and chicken, we determined their ability to dictate time, and location, and quantity of reporter gene expression by employing a transgenic mouse assay. Interestingly, they functioned as distinct *cis*-acting activators, able to direct temporally and spatially specific expression during mouse embryogenesis largely in a non-redundant fashion, mimicking endogenous *Gli3* expression.

Mouse *Gli3* is involved in multitude of patterning mechanisms during early embryonic development, and similarly, its expression pattern is highly complex and well defined. *Gli3* is strongly expressed dorsally in the entire brain region and also in cell populations within ventral aspects of brain, and it is known to play a key role not only in the dorsal-ventral patterning but also anterior posterior patterning of telencephalon, diencephalon, midbrain, and hindbrain (Aoto et al. 2002; Kuschel et al. 2003; Matise et al. 1998; Tyurina et al. 2005). In the spinal cord, the Gli3 functions are required for normal MN (motor neuron) differentiation, for the correct spatial patterning of V0-V2 interneurons (intermediate spinal cord), and for the development of floor plate cells and V3 interneurons (ventral spinal cord) (Bai et al. 2004). Interestingly, although *Gli3* is extensively expressed in the dorsal spinal cord, no obvious phenotype is seen there in *Gli3* mutants (Bai et al. 2004). Mouse *Gli3* is extensively expressed in the developing facial mesenchyme and has critical roles in the medial nasal process, lateral nasal process, maxillary process, palatal and tooth development (Aoto et al. 2002; Hardcastle et al. 1998; Mo et al. 1997). Furthermore, *Gli3* functions are also required for the normal development of ear (Hui and Joyner 1993). *Gli3* $^{-/-}$ embryos exhibit a variety of eye abnormalities ranging from microphthalmia to the absence of any remnant of eye tissue (Franz and Besecke 1991; Furimsky and Wallace 2006; Johnson 1967; Tyurina et al. 2005). The protein is expressed in the stalk region of the optic field, in the neural retina, retinal pigment epithelium, lens and surface ectoderm (Aoto et al. 2002; Zaki et al. 2006). Within the

somite, *Gli3* expression is initially widespread and becomes restricted rapidly to the dorsal medial lip of the dermamyotome (DML, the precursor cells for epaxial musculature) and the ventral lateral lip of the dermamyotome (VLL, the progenitor cells for hypaxial muscles), playing a critical role in epaxial and hypaxial myotome formation (McDermott et al. 2005). *Gli2Gli3* double mutants display a variety of foregut defects from oesophageal atresia, tracheo-oesophageal fistula, severe lung phenotype, smaller than normal hepatic and pancreatic buds, to complete absence of lung, trachea and oesophagus (Motoyama et al. 1998). Consistent with the genetic studies in mice, in some Pallister-Hall syndrome patients, foregut malformations including lung lobulation defects, tracheal stenosis and tracheo-oesophageal fistula have been observed (Kang et al. 1997a; Verloes et al. 1995). Furthermore, *Gli3* has also been implicated in the normal development of stomach (Kim et al. 2005), another foregut derivate. These data suggest key roles of *Gli3* in the normal development of several foregut derivates. Human GLI3 has also been shown to play a role in the normal development of pituitary gland (Kang et al. 1997a). *Gli3* mutations are found to cause renal dysplasia /aplasia in man and mice (Bose et al. 2002; Kang et al. 1997a), and similarly, its mRNA has been detected in the embryonic urogenital structures (Hu et al. 2006). Mouse Gli3 has also been shown to perform a critical role in development of external genitalia and is broadly expressed there (Haraguchi et al. 2001). A genetic study in mice has revealed the importance of Gli3 functions in the normal development of mammary line and mammary placodes (Veltmaat et al. 2006), and by immunohistochemistry the Gli3 protein was localized in all mammary buds (Hatsell and Cowin 2006). The *GLI3* associated dominant genetic disorder Greig cephalopolysyndactyly syndrome (GCPS) and the mouse mutant *extra-toes* (*Xt)* have shown a crucial role of this gene in limb development (Hui and Joyner 1993; Vortkamp et al. 1991). Subsequent studies in mice revealed multiple functions of *Gli3* in limb patterning and morphogenesis for instance:

I) *Gli3* has been implicated in the correct positioning of the limb along the main body axis through a genetic interplay between *dHand* and *Tbx3* (Rallis et al. 2005).

II) Genetic antagonism between *Gli3* and *Shh* within a nascent limb bud regulates the anteroposterior patterning of the limb forming region and of autopod morphology by constraining the digit number and identity (Niswander 2003).

III) One of the typical symptoms of Greig syndrome is severe polysyndactyly. Similarly the limbs of *Gli3* $^{-/-}$ embryos shows severe polysyndactyly, suggesting that along with other functions in limb development, *Gli3* is required for proper apoptosis in interdigital regions.

IV) Recently, it has been shown that a cooperative role of Gli3 and another transcription factor, Plzf (promyelocytic leukaemia zinc finger), is required specifically at the very early stages of limb development for the regulation of cartilage condensations within proximal (stylopod segment) and intermediate elements (zeugopod segment) (Barna et al. 2005). This function of Gli3 is independent of *Shh,* and its role in normal anteroposterior patterning of distal limb elements, as the autopod skeleton was unperturbed in *Gli3*$^{-/-}$ *Plzf*$^{-/-}$ double-knockout mice.

The *Gli3* expression patterns within the nascent limb bud are highly dynamic. Initially, *Gli3* is expressed throughout almost the entire mesenchyme of limb forming regions (along the flank) in an anterior to posterior graded manner (Schweitzer et al. 2000). At later stages, the genetic antagonism between *Gli3* and *dHand* results in restriction of the *Gli3* expression domain largely to the anterior limb mesenchyme. The interaction between *Gli3* and *dHand* ultimately positions the zone of polarizing activity region (ZPA) to the posterior of limb bud (te Welscher et al. 2002). Subsequently, *Gli3* is expressed in the interdigital mesenchyme (Mo et al. 1997). This expression pattern is in agreement with limb specific anomalies in *Gli3* mutants, like positioning of limb along rostrocaudal axis, anteroposterior patterning of distal limb elements, i.e. autopod, polysyndactyly phenotypes. As mentioned previously, Gli3 performs obvious functions in proximal and intermediate skeletal elements of limbs at very early stages of development, however, as Gli3 is broadly present throughout the developing limb without a proximo-distal bias in expression, its expression within proximal mesenchymal condensations (cartilage condensations of stylopod and zeugopod elements) can easily be overlooked (Barna et al. 2005). Nevertheless, recently, through Western analysis and in situ hybridization, Gli3 was detected in the cartilage of developing limb elements (Hilton et al. 2005).

Reflecting the complex roles of Gli3 in early mouse embryogenesis, our results indicate that multiple evolutionary conserved *GLI3* associated human *cis*-regulators control highly coordinated *lacZ* reporter gene expression in transgenic mice, mimicking almost the entire known repertoire of endogenous *Gli3* expression (Fig. 4.4).

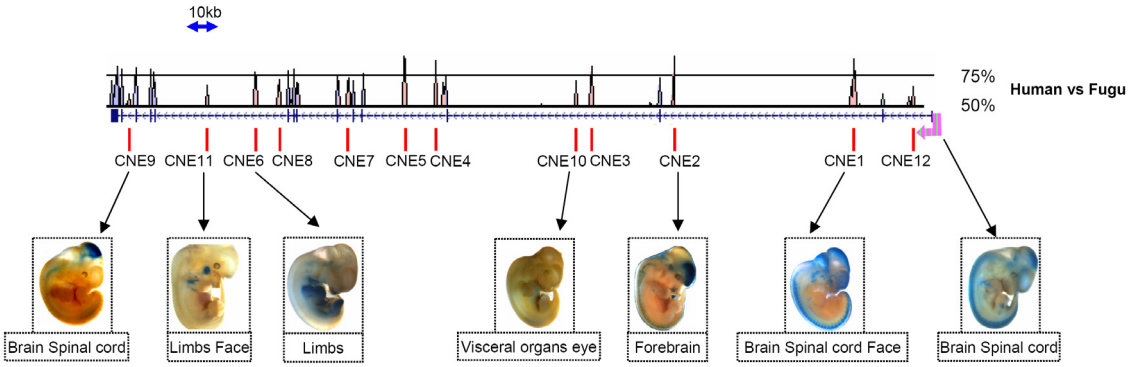


Figure. 4.4. The human *GLI3* associated CNEs act as enhancers, recapitulating almost the entire known repertoire of endogenous *Gli3* expression patterns in mice.

The analyzed human/*Fugu* CNEs are shown by red lines underneath the conservation plot, whereas an upstream 3417 bp non-conserved region encompassing promoter proximal elements is shown by the pink arrow. The direction of the arrow depicts the direction of transcription. For the subset of elements that has been tested in our transgenic mouse assay, the E11.5 embryos are shown along with their primary target sites.

The most obvious activity domain of conserved non-coding element 1 (CNE1) was the developing neural tube (including dorsal midline) where it induced reporter expression from the telencephalon caudally to the tail, in a manner that mimics endogenous *Gli3*-expression pattern (Fig. 3.28 and 3.29). In the brain the CNE1 activity was more obvious in telencephalon and along the dorsal and dorso-lateral portions of diencephalon, mesencephalon, and hindbrain. Additionally, the CNE1 functions were also detected in the ventral neuronal cell populations of the caudal forebrain region as well as the rostral part of hindbrain and medulla oblongata. Within the spinal cord, in addition to roof plate, the CNE1 showed specificity towards intermediate and ventral portions of the spinal card and dictated reporter expression in particular within V0-V3 interneurons, a site where Gli3 functions are well characterized. Consistent with Gli3 roles in the normal development of craniofacial structures and its widespread expression there, the CNE1 was found to drive expression in numbers of facial domains including nasal processes, the derivates of branchial arch material, including maxillary and mandibular components of jaw, Meckel's cartilage and trigeminal (V) ganglion. Furthermore, in the oral cavity, CNE1 recapitulated Gli3 functions in palatal and tooth development. In agreement with the reported Gli3 role in the epaxial, hypaxial, and myotomal compartments of somites and its strong expression within precursor cells for the epaxial and hypaxial muscles, *lacZ* expression governed by CNE1 was observed specifically in the hypaxial buds of inter-limb somites at early developmental stages (e.g. up-to stage E11.5). In addition, the reporter expression was also observed in the lateral dermamyotome-

derived muscle masses (precursors for appendicular skeletal muscles) migrating to forelimbs. No CNE1 activity was observed within the presumptive epaxial compartment of somites. Thus, CNE1 recapitulated a subset of the normal *Gli3* expression patterns within the somites. By E12.5, the CNE1 drove reporter expression strongly within spinal nerves innervating the dorsolateral trunk region and the forelimb.

The reporter expression driven by our previously reported *GLI3* intronic enhancer CNE2 (Paparidis et at 2007) was largely confined to telencephalic portion of forebrain. Whereas the CNE2 mediated *lacZ* activity was present throughout the walls of telencephalic vesicle the CNE1 activity was confined medially. Thus there is a partial overlap in the activities of these two enhancers within the anterior domain of forebrain.

Consistent with Gli3 functions in foregut development, the CNE10 cis-region was found to drive transgene expression in several foregut derivates, which included pharynx, oesophagus, tracheal duct, liver, gall bladder, pancreatic bud, as well as along the wall of stomach and the lumen of duodenum (Fig. 3.30). The CNE10 activity also coincides with Gli3 functions and its expression within embryonic urogenital structures. Gli3 has been shown to be expressed within optic stalk and also within several structures of eye. In agreement with this, at E11.5 CNE10 drove *lacZ* expression in the ventral region of optic stalk and sensory layer of retina. By E12.5, the expression within the eye was more widespread. This element also indicated activity within the presumptive pituitary formatting region (Rathke's pouch) and along the roof of myelencephalon. Although no developmental defect in heart has yet been reported in Gli3 mutants, its mRNA has been localized there (http://www.genecards). Gli3 might be dispensable for heart formation. At E11.5 CNE10 mediated reporter expression specifically and consistently within the ventricular chamber of heart. Consistent with localization of mouse *Gli3* mRNA in all of mammary buds and in the external genitalia, CNE10 induced reporter expression strongly in these tissues.

As discussed previously, the patterning of the anterio-posterior axis of the early limb bud is under the control of two important players of the hedgehog signaling pathway, SHH and GLI3. The key region within the nascent limb, which regulates this patterning, is the extreme posterior territory of the growing limb bud, the ZPA, a signaling center where SHH is expressed. GLI3 is expressed in the limb bud in a domain complementary to SHH. SHH acts as morphogen and controls GLI3 processing in an anterior to posterior graded manner, i.e. a higher ratio of processed repressor form is present anteriorly which decreases towards the posterior part of the developing bud. Whereas a *cis*-acting regulator controlling SHH expression in the ZPA has been defined and well characterized through genetic studies in mice (Lettice et al. 2003), the *cis*-acting region controlling *GLI3* expression domain

complementary to SHH in limb bud was yet unknown. Among the enhancers tested, only CNE6 controlled reporter expression primarily in the growing forelimb and hindlimb bud, in a way that mimicked endogenous GLI3 expression there (Fig. 3.31). At embryonic days 9.5 and 10.5, the CNE6 induced strong reporter expression throughout the limb mesenchyme, except in the extreme posterior margin: a region corresponding the ZPA. In the E11.5 forelimb CNE6 enhanced reporter activity was detected in the proximal and distal regions of limbs, but was absent/weak in the distal posterior region. By E12.5 the activity of this region was detected in the presumptive digits 1-,2-, and 3-forming regions and in the interdigital mesenchyme between digits 1 and 2 as well as between digits 2 and 3. In the posterior half of the handplate the reporter activity was either absent or very weak and thus undetectable. In addition to the limbs, CNE6 directed transgene expression at early stages of development (up-to E10.5) within a rostroventral domain of the telencephalon. By E11.5 the strong *lacZ* activity was seen in the facial region within the precartilage primordium of the nasal septum. This spatial shift in transgene expression from early rostroventral telencephalon to the facial region is consistent with the observation that mesenchymal cells in the latter region are derived from the migration of neural crest cells from the forebrain (Couly et al. 1998; Noden 1983).

Even though both CNE1 and CNE6 directed reporter expression in the nasal system, intriguingly, close histological analysis revealed a separation of their target sites, i.e. CNE1 induced reporter activity specifically within the nasal capsule, whereas CNE6 was active within the nasal septum (Fig. 4.5).

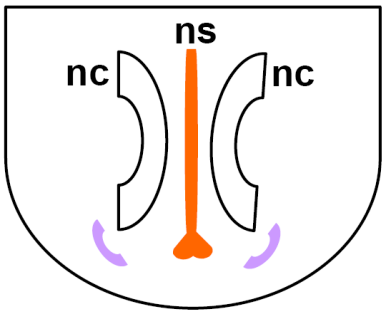


Figure. 4.5. CNE1 and CNE6 activate reporter expression within distinct domains of the nasal system.

Schematic representation of CNE1-directed reporter expression in the pre-cartilaginous nasal capsule (lavender) and CNE6-governed expression specifically within the pre-cartilage primordium of the nasal septum (orange). ns; nasal septum, nc; nasal cavity.

The fore- and hindlimb buds appear at E9 in the future lower cervical/upper thoracic and lower lumber/upper sacral regions respectively as a mass of mesenchymal cells encased in an ectodermal jacket. By E11.5, the limb buds are clearly divided into proximal and distal elements, e.g. future upper arm and handplate in the case of forelimbs. At E12, precartilage condensations of mesenchyme form within limb and are largely replaced by cartilage at about

E13.5 (Kaufman and B.L.Bard 1999). Within the developing limb bud, Gli3 is known to perform two distinct functions: Shh dependent anterior-posterior patterning of handplate and Shh independent proximal cartilage condensations up-to the level of intermediate limb element (zeugopod). The CNE11 activity profile accurately reflects Gli3 positive roles in normal development of proximal and intermediate limb elements (Fig. 3.32). At E11.5, CNE11 governed transgene expression only in the facial region within facial (VII) ganglion. No β-galactosidase staining was observed in the limb buds at this time point. By E12.5, CNE11 regulatory activity was detected in precartilage condensations of mesenchyme within proximal (stylopod) and intermediate elements (zeugopod) of fore and hindlimbs; the *lacZ* expression was excluded from hand and footplates. At E13.5, strong transgene expression was retained within the proximal and intermediate cartilage elements. Within the autopod, the *lacZ* expression was restricted to wrist/ankle and hand/foot elements and was excluded from digits. However, the CNE11 region did not activate transgene expression in the non-cartilaginous mesenchyme encasing the stylopod and zeugopod and in the digital interzones. Thus, in general CNE11 functions were conserved from fish to mouse with respect to its activity within vertebrate appendicular structures. The comparison of transgene expression data from mouse and chick further suggests that time and space for the activity of this enhancer remained highly similar during the course of tetrapod evolution, as in both cases within the limb bud the activity of this enhancer was late in time and was restricted to proximal regions.

The primary domains of CNE9 activity were brain and spinal cord (Fig. 3.33). Unlike CNE1 which governed reporter expression within the entire rostral-caudal axis of the early neural tube, CNE9 activity was limited to midbrain, ventral hindbrain and the ventral spinal cord up-to the level of forelimbs. CNE9 was unable to direct transgene expression in the dorsal midline of the midbrain, however, within the remaining alar plate region of the midbrain this element did induce *lacZ* expression, however, only within the marginal layer. In the ventral midline of the midbrain, the CNE9 activity was more widespread, occupying all three layers (marginal, mantle and ependymal) of neural tissue. When compared to CNE9, the CNE1 drove transgene expression in all three concentric layers of neuronal tissue within dorsal midline and dorso- lateral aspects of the alar plate of midbrain, whereas in the basal plate region it showed activity only within the marginal layer. However, the activity of this element completely vanished in the ventral midline (Fig. 4.6B). Thus CNE1 and CNE9 seem to have overlapping (redundant) activities within the dorso-lateral marginal neuronal tissue of midbrain (green shaded area in Fig. 4.6B). However, they have non-redundant functions along the dorsal and ventral midline, i.e. CNE1 induced transgene expression was more widespread dorsally (including the dorsal midline), while the CNE9 activity was widespread ventrally

(including the ventral midline) and was excluded from dorsal midline (lavender and red colored areas in Fig. 4.6B). Furthermore, CNE1 and CNE9 showed discrete activities in the spinal cord, as CNE1 triggered transgene expression specifically in the dorsal (roofplate), intermediate (V0-V2 interneurons) and ventral spinal cord (V3 interneurons), whereas the CNE9 induced transgene expression was confined to neuronal subpopulations within ventro-lateral (motor neurons) aspects of spinal cord (Fig. 4.6A). In addition to brain and spinal cord, the CNE9 mediated transgene expression in the inter-limb somites within progenitor cells for hypaxial muscles.

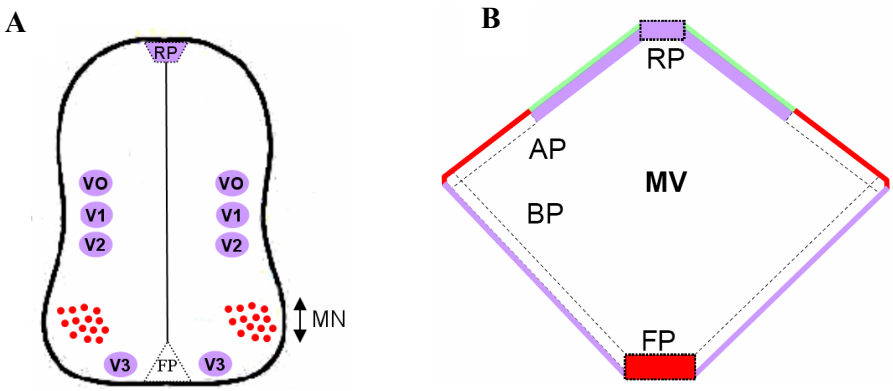


Figure. 4.6. Model of the separate CNE1 and CNE9 activity domains

(A) Spinal cord, (B) midbrain. Lavender colored regions depict the CNE1 activity sites whereas CNE9 target sites are shaded red. Overlapping reporter expression, governed both by CNE1 and 9 is shown by bright green color. In panel B the marginal layer is depicted by the outer thin region. RP, roofplate; FP, floorplate; V, ventral interneurons; MN, motor neurons; AP, alar plate; BP, basal plate; MV, mesencephalic vesicle.

4.8 Activity of a human *GLI3* promoter proximal region

Genes transcribed by RNA polymerase II typically contain two distinct families of *cis*-acting regulatory DNA elements, a promoter and distal regulatory elements, which can be enhancers, silencers, insulators, or locus control regions.

The function of the promoter is to integrate information about the status of the cell in which it resides and to alter the rate of transcription initiation of the associated genes accordingly. A classical vertebrate promoter region is composed of a core promoter element and proximal regulatory elements. The core promoter region could be of ~ 100 bp, whose functions are to provide a docking site for the basic transcriptional machinery and to position the start of transcription relative to the coding sequence (Wray et al. 2003). By itself, the core promoter region is known to activate transcription at a very low level. Moreover, because

most of the transcription factors that bind to the basal promoter are ubiquitously expressed they confer little regulatory specificity (Wray et al. 2003). The proximal promoter is defined as the sequence elements residing immediately up-stream of the core promoter and typically contains binding sites for multiple transcription factors which act synergistically to influence the promoter activity (Maston et al. 2006). Proximal promoter sequences may help selectively to recruit transcription factor complexes positioned at a distant site. Distant enhancers can contact their specific genes through interactions with TFs positioned within the proximal promoter regions (Calhoun et al. 2002). However, the functional distinction between enhancers and the proximal promoter elements is unclear (Maston et al. 2006).

The functional analysis of the human *GLI3* core promoter region has been reported (Paparidis 2005).

In order to access the default *cis*-regulatory capability of *GLI3* proximal promoter elements (without juxtaposition with *GLI3* associated distal enhancers), *in vivo* assays have been performed with a fragment including the first six nucleotides of untranslated exon-1 and extending to +3503 bp upstream of *GLI3* (Fig. 3.17). Consistent with the recently published results in another gene (Fisher et al. 2006) even though the interval span by the ~3.5 kb fragment was not conserved in the teleost lineage, the human element (in combination with the heterologous human *β*-globin promoter) drove reporter gene expression not only in transgenic mice but also in zebrafish embryos (Fig. 3.18 & 3.27). In both species, the target site specifity of *GLI3* proximal promoter region was widespread and largely overlapped with the distant enhancers. For instance, in mice brain and spinal cord the GLI3_PvuII fragment drove transgene expression prominently within a domain overlapping with CNE1 and CNE9 activity sites. The GLI3_ PvuII region was able to induce reporter expression within limbs. However, there the distribution of reporter transcript represented only part of the endogenous *Gli3* expression pattern.

The widespread activity of this interval is attributed to fact that the ~3.5 kb element is large and might encompass numbers of regulatory regions. Further dissection of the region may reveal individual regionally restricted components.

Taken together, functional analysis suggests that, by default the proximal promoter element could mimic part of the aspects of endogenous Gli3 expression. This interval might work in conjunction with distinct enhancer elements at different time or in different tissues to regulate transcription in a spatially / temporally specific manner.

4.9 Two distinct enhancers control *Gli3* expression in the developing limbs

GLI3 functions during early embryonic development define anatomy and morphology of the future limb along two different axes, anterio-posterior (A-P) and proximo-distal (P-D). To pattern the A-P aspects of distal limb elements (autopod), GLI3 works in conjunction with SHH, whereas GLI3 roles in patterning of skeletal elements along the P-D axis up-to the distal margin of the zeugopod are independent of SHH. This study has defined two distinct enhancers apparently recapitulating entirely the known aspects of endogenous *Gli3* expression within cartilaginous and non cartilaginous mesenchyme of embryonic limbs (Fig. 3.31 & 3.32). Furthermore, the target site specificity of these two enhancers is non-redundant (Fig. 4.7) and it closely reflects the two distinct roles of Gli3 in limb patterning and growth. The spatiotemporal activity of CNE6 reflects Gli3 functions in A-P patterning of the autopod (Fig. 4.7). In contrast, with respect to time, CNE11 is a late enhancer and sparks reporter expression specifically within proximal parts (future arm, Fig. 4.7) only when the limb mesenchyme starts to condense to form precartilage (E12), mimicking precisely the Gli3 functions in the patterning of skeletal elements along the upper extremities of the limb (Barna et al. 2005).

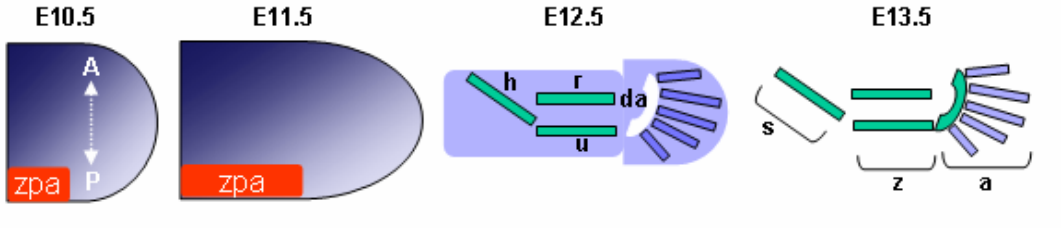


Figure. 4.7. Schematic diagram representing the non-overlapping, limb specific activity domains of CNE6 and CNE11 enhancer regions.

At E10.5 and E11.5, the CNE6 governed reporter actvity was detected throughout the developing limb bud (blue color), with the exception of posterior margin, where Shh is known to be expressed (red block, zpa=zone of polarizing activity). At these time points no CNE11 activity was detected in the limb bud. By E12.5, CNE11 induced transgene expression specifically within precartilage condensations of mesenchyme within presumptive proximal and intermediate limb elements (elements depicted by green bars), whereas it did not show any activity within the handplate precartilage condensations or in the non-cartilaginous mesenchyme encasing the presumptive limb skeleton. At this time point the CNE6 activity (depicted by blue color) was confined to digital rays, digital interzones and to the non-cartilaginous mesenchyme encasing the precartilage condensations of proximal and intermediate limb elements. By E13.5, when the precartilage condensations of mesenchyme are replaced by cartilage, the CNE11 activity (depicted as green colored regions) was not only retained in the stylopod and zeugopod, but was extended more distally up-to cartilaginous elements of digit arch (wrist/ankle). At this time point the CNE6 directed reporter expression can be seen in the individual digits. h, humerus; r, radius; u, ulna; da, digit arch; s, stylopod; z, zeugopod; a, autopod.

Interestingly, when tested in zebrafish both of these human-fish non-coding conserved regions functioned as enhancers, however only CNE11 could evoke reporter expression in the developing pectoral fin bud (homologous of mouse limb bud) (Fig. 3.23). It is possible that failure to detect CNE6 activity within the fish fin bud could be due to insufficient sensitivity of the assay that has been employed (transient reporter gene expression). Nonetheless, the possibility remains that CNE6 might not be involved in the *gli3* expression within the fish pectoral fin bud. Here we suggest, even though CNE6 is an anciently conserved enhancer shared among fish and tetrapods, it might have been co-opted during the course of evolution to regulate *Gli3* expression within more modern aspects of vertebrate appendicular structure, i.e. within hands and feet (autopodia). In contrast, with respect to fin/limb specificity, the CNE11 region largely preserved its ancient functions, dictating GLI3 expression within ancient fin/limb domains, i.e. stylopod and zeugopod. This notion is impelled by following observations:

The various domains of tetrapod limbs and the genetic programs associated with the development of these domains have been shown to be assembled over evolutionary time in the fins of fish. The distal limb elements (autopodia) were previously considered as tetrapod innovation (Shubin et al. 1997). However, a recently discovered member of an extinct sister group of tetrapods, *Tiktaalik roseae* (a link between fishes and land vertebrates) provides evidence that the features of distal limb elements of basal tetrapods were already present in the fins of fish before terrestrial invasion (Shubin et al. 2006). Similarly, the genetic patterns involved in the development of tetrapod autopod have also been observed in fish fin. The posteriorly restricted expression of *Shh* within early limb bud of mouse is a key determinant of autopod patterning along the A-P axis. A comparable domain of *Shh* expression was found in the fins of teleosts (Tyurina et al. 2005), chondrichthyans (Dahn et al. 2007), and in basal actinopterygians (Davis et al. 2007). The occurrence of two spatiotemporally distinct phases of 5´HoxD genes expression, i.e. an early collinear fashion (determining the skeletal morphology of proximal limb elements) and a late–phase with inverted collinear expression (determining the skeletal morphology of the autopod), are correlated with distinct genetic and evolutionary basis of proximal and distal limb regions of tetrapods. Intriguingly, both of these phases of 5´HoxD gene expression were found to be conserved in fish (Davis et al. 2007; Freitas et al. 2007). Despite the fact that both palaeontological and comparative developmental data advocate that autopodia are not an evolutionary novelty of the tetrapod limb, there remain large morphological differences, even between the distal fin bones of *Tiktaalik roseae* and distal limb bones (wrist, ankle and digits) of tetrapods (Ahlberg and

Clack 2006). If the unique features of the distal limb elements in tetrapods evolved from distal fin bones of fish, the process must have involved considerable developmental repatterning.

One possible way to explore the genetic basis of morphological differences among the distal limb elements between fish fin and the tetrapod limb is to elucidate those genes which are common to both fish and tetrapod appendage development, however, expressed differently during early stages of fin and limb development. *Gli3* is among those evolutionary conserved appendicular regulators whose expression dynamics differs among fish fin and tetrapod limb. Genetic studies with mice and chicken have shown that during the early phase of tetrapod limb development the *Gli3* transcript is largely localized to the anterior half, through *Shh* and *dHand* mediated constraints, which themselves are expressed only in the posterior margin of nascent limb bud. Thus during early phases of tetrapod limb development, the *Shh/dHand* and *Gli3* expression domains are non-overlapping. Genetic studies with teleosts (zebrafish) and basal actinopterygia (*Polydon spathula*) have shown that during early stages, *gli3* is expressed uniformly throughout the fin bud without anterior-posterior bias, with expression in the zone of polarizing activity overlapping with *shh* (Davis et al. 2007; Tyurina et al. 2005). At later stages, the *gli3* expression is largely confined to the posterior-distal margin of the fin bud, visually overlapping with *shh* and *dHand* transcripts (Davis et al. 2007). Thus it appears that *gli3* functions and expression territories within the fin bud are *shh* independent, whereas in the limb bud *Gli3* and *Shh* are known to influence each other in expression and function. This model of the role of GLI3 in fin/limb development is summarized in Fig. 4.8. The differences in the expression pattern indicate that *Gli3* may have a distinct cis-regulatory control for fish fin and tetrapod limb. On the other hand, the similar posteriorly restricted localization of the Shh transcript (Fig. 4.8) suggests the occurrence of a conserved mechanism to regulate its expression in fin and limb. In fact, a single anciently conserved *cis*- acting regulator is known to recapitulate the entire aspects of Shh expression in the limb bud, and the functions of this element are conserved among tetrapod and fish lineages (Lettice et al. 2003).

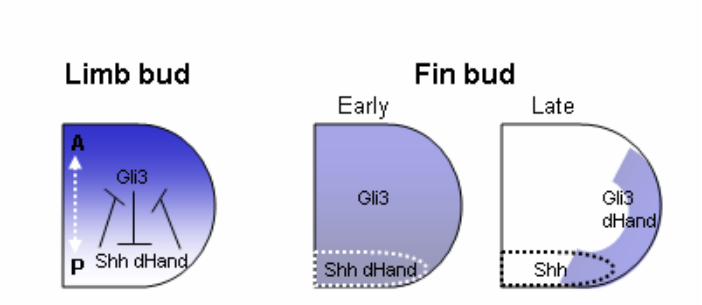


Figure. 4.8. Schematics showing the differences in the expression pattern of *Gli3* during fin and limb bud development.

During early phases of tetrapod limb development the *Shh/dHand* and *Gli3* expression domains (blue) are non-overlapping. In contrast, during early stages of fin development, *gli3* is expressed uniformly throughout the fin bud without anterior-posterior bias, with expression in the zone of polarizing activity (indiacted by dotted area) overlapping with *shh.* At later stages, the *gli3* expression is largely confined to posterior-distal margin of fin bud, visually overlapping with *shh* and *dHand* transcripts.

In this study it has been shown that genetic regulatory mechanisms to control the *Gli3* expression in the tetrapod limbs are fairly complex when compared to Shh, and in accord with its evolving role to govern a more complex morphology they have significantly diverged among tetrapods and fish. We propose that Shh independent functions of Gli3 within proximal skeletal components of mouse limb (stylopod and zeugopod which are comparable to skeletal elements of the fish fin) (Barna et al. 2005) are reminiscent of its ancient expression pattern within the fish fin bud and are under the control of a proximal limb/fin specific enhancer element (CNE11). In contrast the Shh dependent expression and functions of *Gli3* along the distal margin of tetrapod limbs are under the regulation of a distinct enhancer (CNE6). The distal limb morphology defined by *Gli3*, such as digit number and identity have not been determined with certainty in the sister group of limbed vertebrates, *Tiktaalik* (Shubin et al. 2006). Therefore, CNE6-triggered co-option of *Gli3* for the re-patterning of distal limb skeleton might have occurred subsequent to terrestrial invasion in limbed vertebrates (tetrapods) as outlined in Fig. 4.9. The proposal that CNE6 might have gained novel functions during tetrapod evolution is further strengthened by the fact that excision of the entire fish conserved track (179bp) did not abolish the enhancer functions of the human element, as the flanking tetrapod specific conserved intervals were able to up-regulate reporter expression to a significant level in human cell cultures (Fig.3.16). These observations support the notion that, cooption and functional modification of ancestral cis-regulatory sequences have been important in the evolution of morphological traits (Prud'homme et al. 2007; Wray 2007).

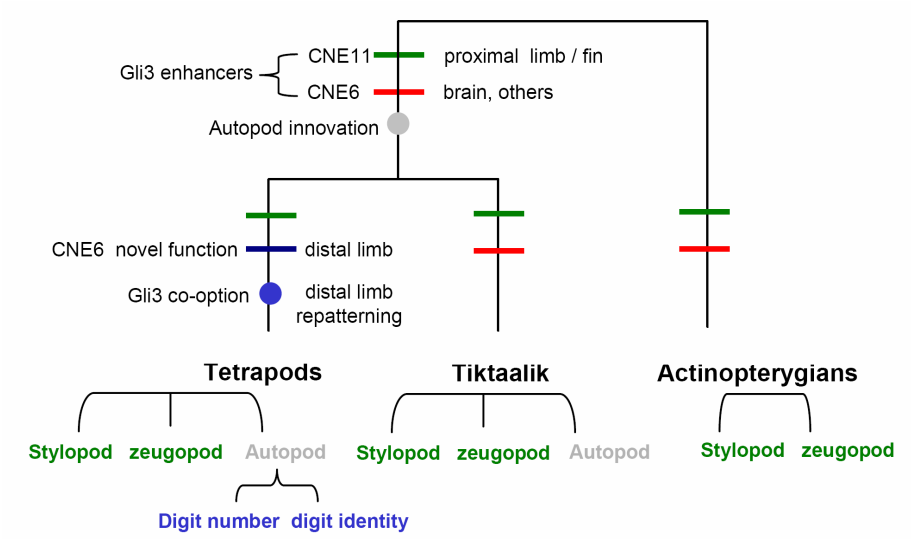


Figure. 4.9. Diagram depicting the functional conservation and divergence of Gli3-associated limb specific cis-regulatory regions among tetrapods, actinopterygians and the extinct sister group of tetrapods, Tiktaalik.

Functional data from zebrafish, mice and chick show that CNE11 functions with respect to limb/fin specificity were largely conserved during the course of evolution (green bar). In contrast, CNE6 might have evolved novel target site specificity (blue bar) to co-opt the *Gli3* gene (blue circle) for the re-patterning of distal limb elements in tetrapods. The innovation of autopod with its basic architecture/morphology might itself be a *Gli3* independent phenomenon (gray circle).

4.10 Filling in the gap: the crosstalk among limb specific cis-regulators

The genetic studies with mice and other model organisms have shown that *Hoxd* genes, *Shh*, *dHand*, and *Gli3* are key regulators of developmental patterning of early limb bud along the A-P axis. The *Gli3* imposed constraints on the posterior restriction of *dHand*, 5´*Hoxd* (d13-10), and *Shh* transcripts, the positive feedback loop among *dHand*, 5´*Hoxd* (*Hoxd*13-*Hoxd*10), and *Shh* as well as their mutual effect on the localization of *Gli3* anteriorly (Litingtung et al. 2002; te Welscher et al. 2002), suggest that products of these genes may influence the transcription of each other within the nascent limb bud through direct or indirect interactions with the relevant *cis*-regulatory sequences. The *cis*-regulatory regions positioned at the 5´ end of the *Hoxd* genes cluster, the GCR (*Global control region*) and the POST (*posterior restriction region*), and the one at the 3´-end of the *Hoxd* cluster, ELCR (*early limb control region*), are known to activate 5´*Hoxd* transcription posteriorly and repress them anteriorly (Tarchini and Duboule 2006; Zhang and Nei 1996). For the *Shh* gene, a single dual nature enhancer element, ZRS (*zone of polarizing activity regulatory sequence*) has been shown to activate and repress its promoter along the posterior and anterior half of the growing

limb bud, respectively (Sagai et al. 2005) (Fig. 4.10). Here, we have identified two distinct limb specific *cis*-acting regulators for the *Gli3* gene. Among them, CNE6 defines *Gli3* expression specifically along the A-P axis of the developing limb bud. The data from cell line assays suggests that this enhancer is dual in nature (activator/repressor), and it is highly likely that it might behave similarly in the growing limb bud to activate or repress *Gli3* expression forming an anterior to posterior gradient through interactions with a differential subset of transcription factors (Fig. 4.10).

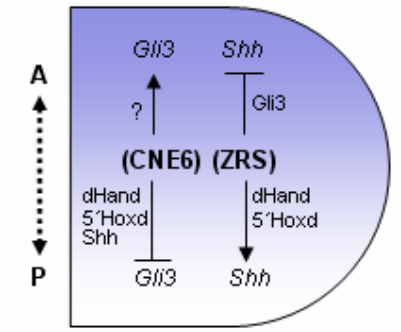


Figure. 4.10. Model depicting a crosstalk among the *Gli3* (CNE6) and *Shh* (ZRS) associated limb specific *cis*-regulators during early developmental patterning of limb along A-P axis

The downstream effectors of Shh signaling may repress *Gli3* transcription posteriorly through interaction with their binding sites within CNE6. Furthermore the posteriorly restricted transcription factors, dHand and 5´Hoxd gene products may influence *Gli3* transcription negatively and *Shh* transcription positively through direct interaction with CNE6 and ZRS, respectively. At the anterior half of the early limb bud, Gli3 probably represses *Shh* transcription directly, through occupancy of its binding sites within the ZRS, whereas its difficult to predict those developmental regulators which might influence *Gli3* transcription anteriorly, as to date no gene is known that could act upstream of *Gli3* to control early patterning of limb along the A-P axis.

We propose that, this crosstalk among Gli3, Shh, 5´Hoxd (d13-10) and dHand, might occur through their limb specific cis-regulators, to correctly partition their activity domains within the growing limb bud. This spatiotemporal partitioning of Gli3, Shh, 5´Hoxd (d13-d10) and dHand transcripts is translated into precise patterning of limbs along A-P axis. A putative crosstalk among limb patterning regulators through their enhancer elements is compiled in Fig. 1.18.

Nonetheless, the possibility remains, that the products of these limb patterning genes might interact directly with each other through protein-protein interactions to counteract or reinforce their respective functions, in particular in the limb domain.

4.11 Multiple independently acting regulatory sequences signal the occurrence of higher levels of modularity in the body plans of modern vertebrates

It has widely been accepted that differences in morphological and anatomical traits among closely related species are correlated to changes in *cis*-acting sequences (Carroll 2005). It now

emerges that *cis*-acting regulatory networks of early developmental regulators are often modular, with multiple independent enhancers mediating the expression of an associated gene in multiple embryonic compartments, independently. Functional changes in one specific *cis*-regulator through mutations might alter the space and time distribution of the associated gene product in one developmental domain, whereas the rest of the expression pattern and the protein activity will largely remain un-interrupted. Thus changes in *cis*-acting sequences will have minimal cost on overall fitness and can serve as raw material for the evolution of morphological and anatomical diversification within and between species.

In this study, we elucidate multiple anciently conserved *cis*-regulatory modules for the evolutionary conserved developmental regulator gene *GLI3*. The functionality of these human elements was tested in three different model organisms (zebrafish, mouse and chick), separated from each other by extreme phylogenetic distance, and also in human cell lines. The most important and unique observation was, that each of these human sequences exhibited the enhancer functions when tested in each of the model organisms or in the human cell lines, and thus preserved their gene regulatory potentials over the course of millions of years of parallel evolutionary divergence time between human, mouse, chick and zebrafish.

Even though the regulatory capability of these human elements was conserved from fish to man, comparative sequence analysis and *in vitro* deletion analysis suggest, that each of these elements might have undergone substantial functional divergence in each lineage. Seven of these intra-*GLI3* CNEs, tested independently in zebrafish, not only recapitulated the known expression repertoire of zebrafish *gli3* but also showed considerable functional redundancy with respect to the site of expression. In contrast, for the same CNEs, the functional data from mice suggest that multiple independent enhancers control the expression of *Gli3* in distinct developmental domains, largely in a non-redundant fashion. In particular, the comparative transgene expression data from mice for CNE1/CNE6 (Fig. 4.5), CNE1/CNE9 (Fig. 4.6) and CNE11/CNE6 (Fig. 4.7) suggests, that distinct cis-regulators may not only dictate the expression of the associated developmental regulator in two discrete developmental sites, but also they can act in a non-redundant manner within the same developmental compartment acting upon distinct cell types. This fine segregation of enhancer functions signals the occurrence of higher levels of modularity among regulatory elements governing the body plans of modern vertebrates (e.g., mammals). Not only adjacent developmental domains but also the cellular subpopulations within same developmental domain are semiautonomous units with respect to expression control of the same developmental regulator. The high level of modularity could also illustrate the strikingly limited pleiotropic effects of mutations affecting the genetic regulatory systems of developmentally important genes in modern

vertebrates (e.g. mammals). This notion underscores the important contribution of *cis*-regulatory sequences for morphological and anatomical evolution and diversification among modern vertebrates.

4.12 Intra-*GLI3* enhancers depict the preservation and divergence of target site specificity during the course of evolution

The comparative mouse-fish data showed that some of these ancient enhancers diverged with respect to target embryonic domains in which they dictate expression in either group, whereas others retained the specificity of their action, at least in part. For instance, in zebrafish the CNE1 directed reporter expression primarily in neuronal subpopulations in brain and spinal cord, and similarly, this element was highly active throughout the neural tube in mice. The forebrain specific regulatory functions of ultraconserved element CNE2 appeared to be exceptionally conserved among fish and tetrapod lineages (Paparidis et al. 2007).

The most prominent sites of CNE10 activity in zebrafish were various subdivisions of eye. In addition it also induced reporter expression significantly in organs like the pericardial region and in the lower jaw primordia. In mice, the functions of the same element seem to be conserved with respect to eye and heart, however CNE10 was unable to show any activity in the mandibular arch region of the mouse embryos. Furthermore CNE10 governed reporter expression in several foregut derivates and in the urogenital structures of developing mouse embryos, whereas no reporter expression was observed in the comparable structures of zebrafish embryos.

In zebrafish, CNE11 directed reporter expression strongly in the developing heart chambers, pectoral fins, and in the muscle fibers. When tested in mice and chick, the primitive functions of this element were found to be preserved with respect to limb/fin specificity throughout the course of evolution; however CNE11 did not show any activity within muscles or heart chambers of transgenic mouse embryos.

CNE6 mediated reporter expression most prominently in the spinal cord and hindbrain neurons of zebrafish embryos and also in the muscle fibers, whereas in mice the enhancer activity of CNE6 was largely confined to fore and hindlimb buds and to the rostroventral telencephalon. Thus, with respect to the site of action, the functions of CNE6 seem to have diverged considerably among fish and mammalian lineages.

The most obvious domains for CNE9 activity were notochord and spinal cord of developing zebrafish. However, in transgenic mice, in addition to the upper cervical region of the spinal cord, the CNE9 functions were observed in midbrain, ventral hindbrain, and also in

the interlimb somites. Furthermore, when compared to fish, no transgene expression was observed in the notochord of mouse embryos.

These findings demonstrate that that even though *Gli3* in fish and mammals share multiple evolutionarily conserved non-coding sequence elements serving as *cis*-acting regulators, the functions of this ancient "gene regulatory catalog" might have diverged at two levels: In mammals these *cis*-regulators attained a higher level of functional modularity by abolishing the potential for redundant expression control. Secondly, in order to cope with differential developmental and anatomical needs of fish and mammals the target site specificity of most of these elements has diverged significantly among these two lineages. This sort of functional differentiation might have been achieved either through changes in the over all span of enhancers or through turnover of transcriptional factor binding site inputs.

4.13 Evolutionary patterns of GLI sequences within and between species

Inspection of average Ka and Ks values (Table 3.3 & 3.4) revealed three important aspects of *GLI* evolutionary patterns. Firstly, all the three *GLI* gene family members showed a significantly higher rate of both silent and non-silent substitutions in fish when compared to mammals, suggesting a relatively relaxed selection in the fish lineage. This pattern correlates well to observations that genes evolve faster in fish than in mammals (Robinson-Rechavi and Laudet 2001). Secondly, between mammalian-fish lineages, the significantly higher average Ka and Ks values for *GLI1* compared to *GLI2* and *GLI3* indicates relaxed selection and accelerated evolution in *GLI1*. This is perhaps reflected in the divergent *GLI1* functions attained in teleosts and tetrapods (Karlstrom et al. 2003), since they last shared a common ancestor 450 Mya. Thirdly, between mammalian-fish *GLI2* and *GLI3* genes, not only the average Ka values (usually subject to selective pressure) but also the corresponding Ks values (assumed to be neutral) are significantly lower than saturation level (Ks >5) (Table 3.4). This indicates that strong purifying selection operates on both silent and non-silent sites. The lower rate of substitutions at silent sites is suggestive of codon usage bias in these two genes (Hellmann et al. 2003; Zhang et al. 2003). Furthermore, average Ka values for *GLI2* and *GLI3* between mammalian-fish lineages are similar, perhaps due to equivalent functional constraints imposed on both genes.

Whilst *GLI1* appears to have undergone rapid evolution since the divergence of tetrapods and teleosts, the *GLI2* and *GLI3* sequences appear to have evolved at considerably slower rate. This data is consistent with the functional conservation of *GLI3* in vertebrates (Tyurina et al. 2005), but not with experimental data that indicates a functional divergence of *GLI2* orthologs in mice and zebrafish (Karlstrom et al. 2003). This functional divergence of *GLI2*

can be explained by two scenarios, by accommodating subtle changes (non-silent) within critical functional domains of the protein in each lineage, leading to functional divergence, or perhaps by changes in gene expression pattern, while keeping the protein activity domains conserved throughout the course of evolution.

The comparison of evolutionary distance between *GLI* paralogs (Table 3.5) in each animal revealed a markedly increased evolutionary rate ($p < 0.05$) of *GLI1* in human and mouse. This may reflect changes in the functions of this gene compared its paralogs in these species, as evident from functional studies, where GLI2 and GLI3 are found to depict overlapping activities in mammalian cell culture and transgenic experiments, while GLI1 appears to perform divergent functions (Lee et al. 1997; Ruiz i Altaba 1998; Ruiz i Altaba 1999; Sasaki et al. 1999; Shin et al. 1999).

4.14 Evolutionary history of map position of the *GLI* paralogs

During the evolutionary history of life on Earth there has been a trend towards drastic transitions from simple to more complex life forms, like from unicellular bacterium to simple multicellular placozans, diploblastic organisms with two germ layers to bilaterians with a third germ layer, simple chordates to vertebrates (Carroll 2001). The innovation of new structures and functions during these macroevolutionary events has in part been accomplished through expansion in the genetic toolkit, e.g. by gene duplications (Ding et al. 2006). In fact, extensive gene duplications have been suggested at the base of vertebrate lineage which resulted in widespread existence of gene families in modern vertebrates (Ding et al. 2006; Gu et al. 2002; McLysaght et al. 2002; Panopoulou et al. 2003; Vandepoele et al. 2004). Expansions in gene number are associated with the evolution of increased morphological and anatomical complexity and diversity achieved by vertebrates compared to basal chordates (cephalochordates/tunicates). The organization of paralogous regions (paralogons) in the human and other vertebrate genomes have led to the hypothesis of multiple block duplication events involving large chromosomal segments or even two rounds of whole genome duplication (2R hypothesis) early in the history of vertebrate evolution after their divergence from an amphioxus-like invertebrate ancestor (Holland et al. 1994; Lundin et al. 2003; Ohno 1970; Skrabanek and Wolfe 1998; Wolfe 2001). In contrast to block duplication events, an alternative model of a continuous wave of small-scale gene duplications (involving single genes or chromosomal segments) was suggested to explain the numerous paralogs in vertebrates (Hughes 1998; Hughes 1999; Hughes et al. 2001).

Phylogenetic trees can be used to test the 2R hypothesis (Fig. 4.11). If two rounds of genome duplication occurred, a tree for four vertebrate paralogous genes should exhibit the

topology of the form (AB)(CD), where the first genome duplication produced the common ancestor of the sequences A/B and C/D and the second genome duplication split these two lineages simultaneously. Thus, under the assumption of the 2R hypothesis the neighboring gene families within potentially quadruplicated regions of the human genome should not only show the same but also the specific type of topology (Hughes 1999). Nevertheless many phylogenetic analyses have not yielded a predominance of (AB)(CD) topologies, instead a high proportion of gene families showed an asymmetrical (A)(BCD) tree, in which one of the four paralogs diverged prior to others, contradicting 2R (Hughes 1998; Hughes 1999; Martin 1999).

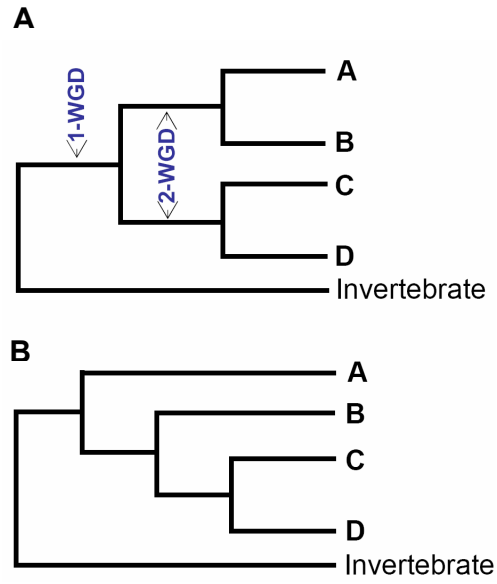


Figure. 4.11. Phylogenetic trees can be used to test the 2R hypothesis.

(A) Tree topology of the type (AB)(CD) for vertebrate gene families having four members supports the occurrence of two rounds of whole genome duplication (WGD) early in the vertebrate history. (B) The topology of the type, where one of the four paralogs diverged prior to others, refutes the 2R assumption.

The four human *HOX* gene clusters bearing chromosomes (Hsa 2, 7, 12, and 17) harbor one of the three large quadrupled genomic regions (paralogon) that have been extensively scrutinized in the literature (Hughes et al. 2001; Larhammar et al. 2002; Lundin et al. 2003; Panopoulou and Poustka 2005; Wolfe 2001). The human *GLI* genes are among the members of the *HOX* cluster paralogon (Fig. 1.20). The fact that two or more paralogs of numerous gene families are linked with *HOX* genes suggests, that these paralogous gene sets along with

the linked *HOX* clusters might have arisen by duplications of an intact chromosomal segment, i.e. through block duplication events. This extensive intra-genomic synteny centered on *HOX* clusters has also been seen as an argument supporting two rounds of whole genome duplication events (2R hypothesis) in the vertebrate lineage (Larhammar et al. 2002; Lundin et al. 2003).

In order to track the evolutionary events involved in structuring the mammalian *HOX*-bearing chromosomes, Hughes and coworkers conducted a phylogenetic analysis of 42 gene families sharing members on two or more of the human chromosomes 2, 7, 12, and 17, the chromosomes that bear *HOX* clusters (Hughes et al. 2001). These authors found that phylogenies of 14 *HOX* linked gene families supported the occurrence of genome duplications before the protostome-deutrostome split. Members of only few families were found to be duplicated within the time window of proposed whole genome/block duplication events. The authors argued that these genes were actually not duplicated simultaneously with the *HOX* clusters because the topologies of their phylogenetic trees were not consistent with the *HOX* cluster phylogeny.

However Larhammar and coworkers (Larhammar et al. 2002) advise caution in rejecting the block/chromosomal duplication hypothesis and argued that only genes that are anciently linked to *H*OX clusters and not those that are transported on the *HOX*-bearing chromosomes as a result of recent rearrangement events should be considered. They recommended the enrichment of sequence information with diverse classes of vertebrates from fishes to mammals to perform more thorough phylogenetic analysis. Larhammar and coworkers concluded that at least 14 gene families on human *HOX*-bearing chromosomes display phylogenetic histories compatible with duplications concomitant with the *HOX* clusters.

4.15 *HOX* linked paralogous regions may not reflect the outcome of ancient block or whole chromosome duplication events

In the present study, the recent accessibility of a huge amount of protein data from sequencing and annotation of increasing numbers of vertebrate genomes was exploited to analyze the phylogenetic history of 11 *HOX* linked gene families (Fig. 1.20 and Table 2.4). The objective was to unravel the constraints that brought the *HOX* clusters and members of multigene families, such as the *GLI* genes, into physical proximity deep in vertebrate history. All of these gene families are anciently linked to *HOX* clusters with 8 families having their members on all human *HOX*-bearing chromosomes, while 3 gene families have paralogs linked to at least three human *HOX* clusters (Fig. 1.20 and Table 2.4). It is of note that 9 of

these families (Table 2.4) are among those 14 gene families, which Larhammar and coworkers (Larhammar et al. 2002) hypothesized to be duplicated simultaneously with the linked *HOX* clusters by block duplication event. For each of these 11 *HOX* linked gene families, the orthologous sequence information from several vertebrate representatives from mammals to bony fishes has been included. Thus, we performed a more robust and thorough phylogenetic analysis compared to previous studies. Given our phylogenetic data, we compared the topologies of those paralogous genes of the each gene families which have arisen within the time window of vertebrates-invertebrates and tetrapods-fishes divergence to test which genes have duplicated concurrently with each other and with the linked *HOX* clusters at the base of vertebrate lineage. We recovered four independent co-duplicated groups involving the members from total 11 gene families. The largest co-duplicated group suggests the simultaneous duplication of members of five gene families (Fig. 4.42A) where the order and close physical linkage of constituent genes is largely disrupted, except the *GLI* and *INHB* genes which are tightly linked on each of the relevant chromosomes (Fig. 1.20). The second co-duplicated group involves members from *ERBB*, *ZNFN1A*, and *IGFBP* families and indicates a conservation of linkage and gene order following co-duplication events (Fig. 4.42B). The *HOX* clusters and members of the *SP* gene family represent the third co-duplicated group (Fig. 4.42C); again the constituent genes remained closely linked on each of the relevant chromosomal segments. The fourth co-duplicated group involves the members from two gene families (Fig. 4.42D), where the linkage between the co-duplicated genes is largely disrupted, except on Hsa17 where *MYL4* is closely linked to the *ITGB3* gene (Fig. 1.20).

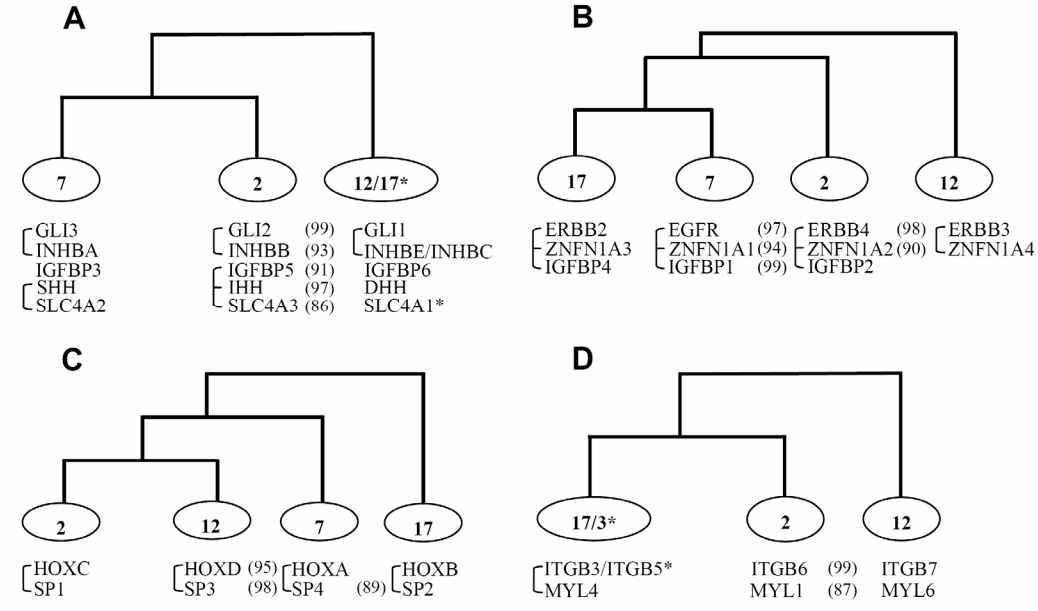


Figure. 4.42. Consistencies in phylogenies of gene families having members on at least three of the HOX-bearing chromosomes.

(A) Schematic topology of *GLI*, *INHB*, *IGFBP*, *HH*, and *SLC4A* families. (B) Schematic topology of *ERBB*, *ZNFN1A* and *IGFBP* family members. (C) Schematic topology of *HOX* clusters and *SP* gene family. (D) Schematic topology of integrin beta chain and myosin light chain gene families. In each case the percentage bootstrap support of the internal branches is given in parentheses. The connecting bars on the left depict the close physical linkage of relevant genes.

Our results show that the extensive triplicate or quadruplicate synteny, that is seen on the present day human *HOX*-bearing chromosomes, is not the outcome of two rounds of duplications experienced by a single ancestral block. Instead, our data suggest that those members of *HOX* linked gene families that arose within the time window of the proposed block duplication events (Fig. 3.40) can be divided into distinct co-duplicated groups. Genes within a particular co-duplicated group share the same evolutionary history and are duplicated in concert with each other, while the genes belonging to different co-duplicated groups may not share the evolutionary history and may not have duplicated simultaneously. We conclude, that gene families with three or more members on human *HOX*-bearing chromosomes might be the outcome of gene-cluster duplication events experienced by vertebrates at different time points in their evolutionary history, whereas their current triplicate or quadruplicate distribution on these chromosomes might be the consequence of chromosomal redistribution of multigene family members through extensive rearrangement of genomic segments encompassing multiple contiguous genes. This would imply that although different co-duplicated groups within human chromosomes 2, 7, 12, and 17 are remnants of waves of

small-scale duplications (segmental/gene-cluster) and chromosomal rearrangement events, they do not indicate a single ancestral block.

This conclusion leaves the question unanswered, if the present co-localization of the members paralogons on the same chromosomes is a result of chance or if they are combined by functional constraints.

The paralogous *GLI* genes, the hedgehog gene family, and the members of *HOX* clusters interact during development, e.g. in the patterning of vertebrates limbs. It is conceivable that long range regulatory mechanisms favour not only a combination of genes in clusters in which the order of the relevant genes is conserved on each chromosome but that a functional advantage favours the collection of functionally related genes on the same chromosomes, irrespective of order and distance.

Abbreviations

AGRN	Ancient gene regulatory network
bp	Base pair
CNS	Central nervous system
ELCR	Early limb control region
GCR	Global control region
GFP	Green Fluorescent Protein
hc	Highly conserved
HEPES	1-Piperazineethane sufonic acid
hpf	Hours post fertilization
kDa	KiloDaltons
LAGAN	Limited area global alignment of nucleotides
Mbp	Millions of base pairs
MN	Motor neurons
Mya	Millions of years ago
PipMaker	Percent identity plot maker
POST	Posterior restriction
PWM	Position weight matrices
RFP	Red Fluorescent Protein
TFBS	Transcriptional factor binding site
VISTA	Visualization tool of alignment
WGD	Whole genome duplication
Wt	Wild type
ZPA	Zone of polarizing activity region
ZRS	Zone of polarizing activity regulatory sequence

REFERENCES

Ahlberg PE, Clack JA (2006) Palaeontology: a firm step from water to land. Nature 440:747-9

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-10

Anand S, Wang WC, Powell DR, Bolanowski SA, Zhang J, Ledje C, Pawashe AB, Amemiya CT, Shashikant CS (2003) Divergence of Hoxc8 early enhancer parallels diverged axial morphologies between mammals and fishes. Proc Natl Acad Sci U S A 100:15666-9

Aoto K, Nishimura T, Eto K, Motoyama J (2002) Mouse GLI3 regulates Fgf8 expression and apoptosis in the developing neural tube, face, and limb bud. Dev Biol 251:320-32

Aoyagi N, Wassarman DA (2000) Genes encoding Drosophila melanogaster RNA polymerase II general transcription factors: diversity in TFIIA and TFIID components contributes to gene-specific transcriptional regulation. J Cell Biol 150:F45-50

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S (1995) Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. Proc Natl Acad Sci U S A 92:1684-8

Bai CB, Stephen D, Joyner AL (2004) All mouse ventral spinal cord patterning by hedgehog is Gli dependent and involves an activator function of Gli3. Dev Cell 6:103-15

Bailey WJ, Kim J, Wagner GP, Ruddle FH (1997) Phylogenetic reconstruction of vertebrate Hox cluster duplications. Mol Biol Evol 14:843-53

Barna M, Pandolfi PP, Niswander L (2005) Gli3 and Plzf cooperate in proximal limb patterning at early stages of limb development. Nature 436:277-81

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321-5

Bejerano G, Siepel AC, Kent WJ, Haussler D (2005) Computational screening of conserved genomic DNA in search of functional noncoding elements. Nat Methods 2:535-45

Boffelli D, Nobrega MA, Rubin EM (2004) Comparative genomics at the vertebrate extremes. Nat Rev Genet 5:456-65

Bose J, Grotewold L, Ruther U (2002) Pallister-Hall syndrome phenotype in mice mutant for Gli3. Hum Mol Genet 11:1129-35

Bray N, Dubchak I, Pachter L (2003) AVID: A global alignment program. Genome Res 13:97-102

Brudno M, Do CB, Cooper GM, Kim MF, Davydov Ej (2003a) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721-31

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S (2003b) Glocal alignment: finding rearrangements during alignment. Bioinformatics 19 Suppl 1:i54-62

Budd G (1996) The morphology of Opabinia regalis and the reconstruction of the arthopod stem-group. Lethaia 29:1-14

Bulyk ML (2003) Computational prediction of transcription-factor binding site locations. Genome Biol 5:201

Calhoun VC, Stathopoulos A, Levine M (2002) Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. Proc Natl Acad Sci U S A 99:9243-7

Campbell K, Flavin N, Ivens A, Robert B, Buckingham M, Williamson R (1989) The human homeobox gene HOX7 maps to 4p16.1 and is deleted in Wolf-Hirschhorn syndrome patients. Am J Hum Genet 45:A179

Canestro C, Bassham S, Postlethwait JH (2003) Seeing chordate evolution through the Ciona genome sequence. Genome Biol 4:208

Carroll SB (2001) Chance and necessity: the evolution of morphological complexity and diversity. Nature 409:1102-9

Carroll SB (2005) Evolution at two levels: on genes and form. PLoS Biol 3:e245

Charite J, McFadden DG, Olson EN (2000) The bHLH transcription factor dHAND controls Sonic hedgehog expression and establishment of the zone of polarizing activity during limb development. Development 127:2461-70

Cisne JW (1974) Evolution of the world fauna of aquatic free-living arthropods. Evolution 28:337-366

Coates MI, Clack JA (1990) Polydactyly in the earliest tetrapod limbs. Nature 347:66-69

Coates MI, Cohn MJ (1998) Fins, limbs, and tails: outgrowth and axial patterning in vertebrate evolution. BioEssays 20:371-381

Conway SJ, Henderson DJ, Copp AJ (1997) Pax3 is required for cardiac neural crest migration in the mouse: evidence from the splotch (Sp2H) mutant. Development 124:505-14

Couly G, Grapin-Botton A, Coltey P, Ruhin B, Le Douarin NM (1998) Determination of the identity of the derivatives of the cephalic neural crest: incompatibility between Hox gene expression and lower jaw development. Development 125:3445-59

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 389:876-81

Dahn RD, Davis MC, Pappano WN, Shubin NH (2007) Sonic hedgehog function in chondrichthyan fins and the evolution of appendage patterning. Nature 445:311-4

Das RM, Van Hateren NJ, Howell GR, Farrell ER, Bangs FK, Porteous VC, Manning EM, McGrew MJ, Ohyama K, Sacco MA, Halley PA, Sang HM, Storey KG, Placzek M, Tickle C, Nair VK, Wilson SA (2006) A robust system for RNA interference in the chicken using a modified microRNA operon. Dev Biol 294:554-63

Davis MC, Dahn RD, Shubin NH (2007) An autopodial-like pattern of Hox expression in the fins of a basal actinopterygian fish. Nature 447:473-6

Davis MC, Shubin NH, Force A (2004) Pectoral fin and girdle development in the basal actinopterygians Polyodon spathula and Acipenser transmontanus. J Morphol 262:608-28

Ding G, Kang J, Liu Q, Shi T, Pei G, Li Y (2006) Insights into the coupling of duplication events and macroevolution from an age profile of animal transmembrane gene families. PLoS Comput Biol 2:e102

Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature 434:857-63

Felsenstein J (1985) Confidence limit on phylogenies: An approach using the bootstrap. Evolution 39:95-105

Fernandez-Teran M, Piedra ME, Kathiriya IS, Srivastava D, Rodriguez-Rey JC, Ros MA (2000) Role of dHAND in the anterior-posterior polarization of the limb bud: implications for the Sonic hedgehog pathway. Development 127:2133-42

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312:276-9

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-45

Franz T, Besecke A (1991) The development of the eye in homozygotes of the mouse mutant Extra-toes. Anat Embryol (Berl) 184:355-61

Freitas R, Zhang G, Cohn MJ (2007) Biphasic Hoxd gene expression in shark paired fins reveals an ancient origin of the distal limb domain. PLoS ONE 2:e754

Furimsky M, Wallace VA (2006) Complementary Gli activity mediates early patterning of the mouse visual system. Dev Dyn 235:594-605

Furlong RF, Holland PW (2004) Polyploidy in vertebrate ancestry: Ohno and beyond. Biological Journal of the Linnean Society 82:425-430

Gerhart J, Lowe C, Kirschner M (2005) Hemichordates and the origin of chordates. Curr Opin Genet Dev 15:461-7

Ghali L, Wong ST, Green J, Tidman N, Quinn AG (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 113:595-9

Gomez-Skarmeta JL, Lenhard B, Becker TS (2006) New technologies, new findings, and new concepts in the study of vertebrate cis-regulatory sequences. Dev Dyn 235:870-85

Goode DK, Snell P, Smith SF, Cooke JE, Elgar G (2005) Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. Genomics 86:172-81

Gottgens B, Barton LM, Gilbert JG, Bench AJ, Sanchez MJ, Bahn S, Mistry S, Grafham D, McMurray A, Vaudin M, Amaya E, Bentley DR, Green AR, Sinclair AM (2000) Analysis of vertebrate SCL loci identifies conserved enhancers. Nat Biotechnol 18:181-6

Grachtchouk M, Mo R, Yu S, Zhang X, Sasaki H, Hui CC, Dlugosz AA (2000) Basal cell carcinomas in mice overexpressing Gli2 in skin. Nat Genet 24:216-7

Grice EA, Rochelle ES, Green ED, Chakravarti A, McCallion AS (2005) Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. Hum Mol Genet 14:3837-45

Gu X, Wang Y, Gu J (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet 31:205-9

Harafuji N, Keys DN, Levine M (2002) Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. Proc Natl Acad Sci U S A 99:6802-5

Haraguchi R, Mo R, Hui C, Motoyama J, Makino S, Shiroishi T, Gaffield W, Yamada G (2001) Unique functions of Sonic hedgehog signaling during external genitalia development. Development 128:4241-50

Hardcastle Z, Mo R, Hui CC, Sharpe PT (1998) The Shh signalling pathway in tooth development: defects in Gli2 and Gli3 mutants. Development 125:2803-11

Hatsell SJ, Cowin P (2006) Gli3-mediated repression of Hedgehog targets is required for normal mammary development. Development 133:3661-70

Hellmann I, Zollner S, Enard W, Ebersberger I, Nickel B, Paabo S (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res 13:831-7

Hersh BM, Carroll SB (2005) Direct regulation of knot gene expression by Ultrabithorax and the evolution of cis-regulatory elements in Drosophila. Development 132:1567-77

Hilton MJ, Tu X, Cook J, Hu H, Long F (2005) Ihh controls cartilage development by antagonizing Gli3, but requires additional effectors to regulate osteoblast and vascular development. Development 132:4339-51

Hinchliffe JR (2002) Developmental basis of limb evolution. Int J Dev Biol 46:835-45

Holland PW, Garcia-Fernandez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. Dev Suppl:125-33

Hu G, Vastardis H, Bendall AJ, Wang Z, Logan M, Zhang H, Nelson C, Stein S, Greenfield N, Seidman CE, Seidman JG, Abate-Shen C (1998) Haploinsufficiency of MSX1: a mechanism for selective tooth agenesis. Mol Cell Biol 18:6044-51

Hu MC, Mo R, Bhella S, Wilson CW, Chuang PT, Hui CC, Rosenblum ND (2006) GLI3-dependent transcriptional repression of Gli1, Gli2 and kidney patterning genes disrupts renal morphogenesis. Development 133:569-78

Hughes AL (1998) Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. Mol Biol Evol 15:854-70

Hughes AL (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. J Mol Evol 48:565-76

Hughes AL, da Silva J, Friedman R (2001) Ancient genome duplications did not structure the human Hox-bearing chromosomes. Genome Res 11:771-80

Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. Mol Biol Evol 10:1360-9

Hui CC, Joyner AL (1993) A mouse model of greig cephalopolysyndactyly syndrome: the extra-toesJ mutation contains an intragenic deletion of the Gli3 gene. Nat Genet 3:241-6

Iotsova V, Caamano J, Loy J, Yang Y, Lewin A, Bravo R (1997) Osteopetrosis in mice lacking NF-kappaB1 and NF-kappaB2. Nat Med 3:1285-9

Ishikawa N, Shimada N, Takagi Y, Ishijima Y, Fukuda M, Kimura N (2003) Molecular evolution of nucleoside diphosphate kinase genes: conserved core structures and multiple-layered regulatory regions. J Bioenerg Biomembr 35:7-18

Jeffery WR, Strickler AG, Yamamoto Y (2004) Migratory neural crest-like cells form body pigmentation in a urochordate embryo. Nature 431:696-9

Johnson DR (1967) Extra-toes: anew mutant gene causing multiple abnormalities in the mouse. J Embryol Exp Morphol 17:543-81

Johnston JJ, Olivos-Glander I, Killoran C, Elson E, Turner JT, Peters KF, Abbott MH, Aughton DJ, Aylsworth AS, Bamshad MJ, Booth C, Curry CJ, David A, Dinulos MB, Flannery DB, Fox MA, Graham JM, Grange DK, Guttmacher AE, Hannibal MC, Henn W, Hennekam RC, Holmes LB, Hoyme HE, Leppig KA, Lin AE, Macleod P, Manchester DK, Marcelis C, Mazzanti L, McCann E, McDonald MT, Mendelsohn NJ, Moeschler JB, Moghaddam B, Neri G, Newbury-Ecob R, Pagon RA, Phillips JA, Sadler LS, Stoler JM, Tilstra D, Walsh Vockley CM, Zackai EH, Zadeh TM, Brueton L, Black GC, Biesecker LG (2005) Molecular and clinical analyses of Greig cephalopolysyndactyly and Pallister-Hall syndromes: robust phenotype prediction from the type and position of GLI3 mutations. Am J Hum Genet 76:609-22

Kalff-Suske M, Wild A, Topp J, Wessling M, Jacobsen EM, Bornholdt D, Engel H, Heuer H, Aalfs CM, Ausems MG, Barone R, Herzog A, Heutink P, Homfray T, Gillessen-Kaesbach G, Konig R, Kunze J, Meinecke P, Muller D, Rizzo R, Strenge S, Superti-Furga A, Grzeschik KH (1999) Point mutations throughout the GLI3 gene cause Greig cephalopolysyndactyly syndrome. Hum Mol Genet 8:1769-77

Kang S, Allen J, Graham JM, Jr., Grebe T, Clericuzio C, Patronas N, Ondrey F, Green E, Schaffer A, Abbott M, Biesecker LG (1997a) Linkage mapping and phenotypic analysis of autosomal dominant Pallister-Hall syndrome. J Med Genet 34:441-6

Kang S, Graham JM, Jr., Olney AH, Biesecker LG (1997b) GLI3 frameshift mutations cause autosomal dominant Pallister-Hall syndrome. Nat Genet 15:266-8

Kanters E, Pasparakis M, Gijbels MJ, Vergouwe MN, Partouns-Hendriks I, Fijneman RJ, Clausen BE, Forster I, Kockx MM, Rajewsky K, Kraal G, Hofker MH, de Winther MP (2003) Inhibition of NF-kappaB activation in macrophages increases atherosclerosis in LDL receptor-deficient mice. J Clin Invest 112:1176-85

Karlstrom RO, Tyurina OV, Kawakami A, Nishioka N, Talbot WS, Sasaki H, Schier AF (2003) Genetic analysis of zebrafish gli1 and gli2 reveals divergent requirements for gli genes in vertebrate development. Development 130:1549-64

Kaufman MH, B.L.Bard J (1999) The anatomical basis of mouse development. Elsevier

Kim JH, Huang Z, Mo R (2005) Gli3 null mice display glandular overgrowth of the developing stomach. Dev Dyn 234:984-91

Kim SK, Selleri L, Lee JS, Zhang AY, Gu X, Jacobs Y, Cleary ML (2002) Pbx1 inactivation disrupts pancreas development and in Ipf1-deficient mice promotes diabetes mellitus. Nat Genet 30:430-5

Kinzler KW, Bigner SH, Bigner DD, Trent JM, Law ML, O'Brien SJ, Wong AJ, Vogelstein B (1987) Identification of an amplified, highly expressed gene in a human glioma. Science 236:70-3

Kinzler KW, Vogelstein B (1990) The GLI gene encodes a nuclear protein which binds specific sequences in the human genome. Mol Cell Biol 10:634-42

Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76:8-32

Kleinjan DJ, van Heyningen V (1998) Position effect in human genetic disease. Hum Mol Genet 7:1611-8

Kmita M, Fraudeau N, Herault Y, Duboule D (2002) Serial deletions and duplications suggest a mechanism for the collinearity of Hoxd genes in limbs. Nature 420:145-50

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3:RESEARCH0008

Kruger G, Gotz J, Kvist U, Dunker H, Erfurth F, Pelz L, Zech L (1989) Greig syndrome in a large kindred due to reciprocal chromosome translocation t(6;7)(q27;p13). Am J Med Genet 32:411-6

Kuschel S, Ruther U, Theil T (2003) A disrupted balance between Bmp/Wnt and Fgf signaling underlies the ventralization of the Gli3 mutant telencephalon. Dev Biol 260:484-95

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Larhammar D, Lundin LG, Hallbook F (2002) The human Hox-bearing chromosome regions did arise by block or chromosome (or even genome) duplications. Genome Res 12:1910-20

Lee J, Platt KA, Censullo P, Ruiz i Altaba A (1997) Gli1 is a target of Sonic hedgehog that induces ventral neural tube development. Development 124:2537-52

Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW (2003) Identification of conserved regulatory elements by comparative genome analysis. J Biol 2:13

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, 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 12:1725-35

Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424:147-51
Lewis PM, Dunn MP, McMahon JA, Logan M, Martin JF, St-Jacques B, McMahon AP (2001) Cholesterol modification of sonic hedgehog is required for long-range signaling activity and effective modulation of signaling by Ptc1. Cell 105:599-612
Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature 409:847-9
Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150-74
Litingtung Y, Chiang C (2000) Specification of ventral neuron types is mediated by an antagonistic interaction between Shh and Gli3. Nat Neurosci 3:979-85
Litingtung Y, Dahn RD, Li Y, Fallon JF, Chiang C (2002) Shh and Gli3 are dispensable for limb skeleton formation but regulate digit number and identity. Nature 418:979-83
Loots GG, Ovcharenko I (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res 32:W217-21
Lundin LG, Larhammar D, Hallbook F (2003) Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates
Manzanares M, Cordes S, Ariza-McNaughton L, Sadl V, Maruthainar K, Barsh G, Krumlauf R (1999) Conserved and distinct roles of kreisler in regulation of the paralogous Hoxa3 and Hoxb3 genes. Development 126:759-69
Margulies EH, Chen CW, Green ED (2006) Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. Trends Genet 22:187-93
Martin AP (1999) Increasing genomic complexity by gene duplication and the origin of vertebrates. Am.Nat 154:111-128
Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29-59
Matise MP, Epstein DJ, Park HL, Platt KA, Joyner AL (1998) Gli2 is required for induction of floor plate and adjacent cells, but not most ventral neurons in the mouse central nervous system. Development 125:2759-70
Matise MP, Joyner AL (1999) Gli genes in development and cancer. Oncogene 18:7852-9
Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34:D108-10
Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16:1046-7
McDermott A, Gustafsson M, Elsam T, Hui CC, Emerson CP, Jr., Borycki AG (2005) Gli2 and Gli3 have redundant and context-dependent function in skeletal muscle formation. Development 132:345-57
McEwen GK, 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 16:451-65
McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. Nat Genet 31:200-4
McShea D (1991) Complexity and evolution: What everybody knows Biol Philosophy 6:303-324
Mercader N, Leonardo E, Azpiazu N, Serrano A, Morata G, Martinez C, Torres M (1999) Conserved regulation of proximodistal limb axis development by Meis1/Hth. Nature 402:425-9

Mo R, Freer AM, Zinyk DL, Crackower MA, Michaud J, Heng HH, Chik KW, Shi XM, Tsui LC, Cheng SH, Joyner AL, Hui C (1997) Specific and redundant functions of Gli2 and Gli3 zinc finger genes in skeletal patterning and development. Development 124:113-23

Moens CB, Fritz A (1999) Techniques in neural development. Methods Cell Biol 59:253-72

Morrison A, Ariza-McNaughton L, Gould A, Featherstone M, Krumlauf R (1997) HOXD4 and regulation of the group 4 paralog genes. Development 124:3135-46

Motoyama J, Liu 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 20:54-7

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 126:2103-16

Nadeau JH, Sankoff D (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147:1259-66

Niswander L (2003) Pattern formation: old models out on a limb. Nat Rev Genet 4:133-43

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. Science 302:413

Noden DM (1983) The role of the neural crest in patterning of avian cranial skeletal, connective, and muscle tissues. Dev Biol 96:144-65

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, Heidelberg

Panman L, Drenth T, Tewelscher P, Zuniga A, Zeller R (2005) Genetic interaction of Gli3 and Alx4 during limb development. Int J Dev Biol 49:443-8

Panopoulou G, Hennig S, Groth D, Krause A, Poustka AJ, Herwig R, Vingron M, Lehrach H (2003) New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res 13:1056-66

Panopoulou G, Poustka AJ (2005) Timing and mechanism of ancient vertebrate genome duplications -- the adventure of a hypothesis. Trends Genet 21:559-67

Paparidis Z (2005) CIS-ACTING ELEMENTS CONTROLLING THE EXPRESSION OF THE HUMAN GLI3 GENE Institute of Human Genetics. Philipps University, Marburg

Paparidis Z, Abbasi AA, Malik S, Goode DK, Callaway H, Elgar G, deGraaff E, Lopez-Rios J, Zeller R, Grzeschik KH (2007) Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. Dev Growth Differ 49:543-53

Park HL, Bai C, Platt KA, Matise MP, Beeghly A, Hui CC, Nakashima M, Joyner AL (2000) Mouse Gli1 mutants are viable but have defects in SHH signaling in combination with a Gli2 mutation. Development 127:1593-605

Peichel CL, Abbott CM, Vogt TF (1996) Genetic and physical mapping of the mouse Ulnaless locus. Genetics 144:1757-67

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499-502

Pohl TM, Mattei MG, Ruther U (1990) Evidence for allelism of the recessive insertional mutation add and the dominant mouse mutation extra-toes (Xt). Development 110:1153-7

Postlethwait J, Amores A, Cresko W, Singer A, Yan YL (2004) Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet 20:481-90

Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA (2005) In vivo characterization of a vertebrate ultraconserved enhancer. Genomics 85:774-81

Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A 104 Suppl 1:8605-12

Radhakrishna U, Bornholdt D, Scott HS, Patel UC, Rossier C, Engel H, Bottani A, Chandal D, Blouin JL, Solanki JV, Grzeschik KH, Antonarakis SE (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 65:645-55

Rallis C, Del Buono J, Logan MP (2005) Tbx3 can alter limb position along the rostrocaudal axis of the developing embryo. Development 132:1961-70

Remenyi A, Scholer HR, Wilmanns M (2004) Combinatorial control of gene expression. Nat Struct Mol Biol 11:812-5

Riddle RD, Johnson RL, Laufer E, Tabin C (1993) Sonic hedgehog mediates the polarizing activity of the ZPA. Cell 75:1401-16

Roberts WM, Douglass EC, Peiper SC, Houghton PJ, Look AT (1989) Amplification of the gli gene in childhood sarcomas. Cancer Res 49:5407-13

Robinson-Rechavi M, Laudet V (2001) Evolutionary rates of duplicate genes in fish and mammals. Mol Biol Evol 18:681-3

Roessler E, Du YZ, Mullor JL, Casas E, Allen WP, Gillessen-Kaesbach G, Roeder ER, Ming JE, 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 100:13424-9

Ruiz i Altaba A (1998) Combinatorial Gli gene function in floor plate and neuronal inductions by Sonic hedgehog. Development 125:2203-12

Ruiz i Altaba A (1999) Gli proteins encode context-dependent positive and negative functions: implications for development and disease. Development 126:3205-16

Ruppert JM, Kinzler KW, Wong AJ, Bigner SH, Kao FT, Law ML, Seuanez HN, O'Brien SJ, Vogelstein B (1988) The GLI-Kruppel family of human genes. Mol Cell Biol 8:3104-13

Ruvinsky I, Silver LM, Gibson-Brown JJ (2000) Phylogenetic analysis of T-Box genes demonstrates the importance of amphioxus for understanding evolution of the vertebrate genome. Genetics 156:1249-57

Ruvkun G, Hobert O (1998) The taxonomy of developmental control in Caenorhabditis elegans. Science 282:2033-41

Sabherwal N, Bangs F, Roth R, Weiss B, Jantz K, Tiecke E, Hinkel GK, Spaich C, Hauffa BP, van der Kamp H, Kapeller J, Tickle C, Rappold G (2007) Long-range conserved non-coding SHOX sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. Hum Mol Genet 16:210-22

Sagai T, Hosoya M, Mizushina Y, Tamura M, Shiroishi T (2005) Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development 132:797-803

Sagai T, Masuya H, Tamura M, Shimizu K, Yada Y, Wakana S, Gondo Y, Noda T, Shiroishi T (2004) Phylogenetic conservation of a limb-specific, cis-acting regulator of Sonic hedgehog (Shh). Mamm Genome 15:23-34

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-25

Sandelin A, Wasserman WW, Lenhard B (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res 32:W249-52

Sanges R, Kalmar E, Claudiani P, D'Amato M, Muller F, Stupka E (2006) Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. Genome Biol 7:R56

Sasaki H, Nishizaki 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 126:3915-24

Scherer SW, Cheung J, MacDonald JR, Osborne LR, Nakabayashi K, Herbrick JA, Carson AR, Parker-Katiraee L, Skaug J, Khaja R, Zhang J, Hudek AK, Li M, Haddad M, Duggan GE, Fernandez BA, Kanematsu E, Gentles S, Christopoulos CC, Choufani S, Kwasnicka D, Zheng XH, Lai Z, Nusskern D, Zhang Q, Gu Z, Lu F, Zeesman S, Nowaczyk MJ, Teshima I, Chitayat D, Shuman C, Weksberg R, Zackai EH, Grebe TA, Cox SR, Kirkpatrick SJ, Rahman N, Friedman JM, Heng HH, Pelicci PG, Lo-Coco F, Belloni E, Shaffer LG, Pober B, Morton CC, Gusella JF, Bruns GA, Korf BR, Quade BJ, Ligon AH, Ferguson H, Higgins AW, Leach NT, Herrick SR, Lemyre E, Farra CG, Kim HG, Summers AM, Gripp KW, Roberts W, Szatmari P, Winsor EJ, Grzeschik KH, Teebi A, Minassian BA, Kere J, Armengol L, Pujana MA, Estivill X, Wilson MD, Koop BF, Tosi S, Moore GE, Boright AP, Zlotorynski E, Kerem B, Kroisel PM, Petek E, Oscier DG, Mould SJ, Dohner H, Dohner K, Rommens JM, Vincent JB, Venter JC, Li PW, Mural RJ, Adams MD, Tsui LC (2003) Human chromosome 7: DNA sequence and biology. Science 300:767-72

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103-7

Schweitzer R, Vogan KJ, Tabin CJ (2000) Similar expression and regulation of Gli2 and Gli3 in the chick limb bud. Mech Dev 98:171-4

Shashikant CS, Kim CB, Borbely MA, Wang WC, Ruddle FH (1998) Comparative studies on mammalian Hoxc8 early enhancer sequence reveal a baleen whale-specific deletion of a cis-acting element. Proc Natl Acad Sci U S A 95:15446-51

Shimeld SM, Holland PW (2000) Vertebrate innovations. Proc Natl Acad Sci U S A 97:4449-52

Shin SH, Kogerman P, Lindstrom E, Toftgard R, Biesecker LG (1999) GLI3 mutations in human disorders mimic Drosophila cubitus interruptus protein functions and localization. Proc Natl Acad Sci U S A 96:2880-4

Shubin N, Tabin C, Carroll S (1997) Fossils, genes and the evolution of animal limbs. Nature 388:639-48

Shubin NH, Daeschler EB, Jenkins FA, Jr. (2006) The pectoral fin of Tiktaalik roseae and the origin of the tetrapod limb. Nature 440:764-71

Simmons AD, Horton S, Abney AL, Johnson JE (2001) Neurogenin2 expression in ventral and dorsal spinal neural tube progenitor cells is regulated by distinct enhancers. Dev Biol 229:327-39

Skrabanek L, Wolfe KH (1998) Eukaryote genome duplication - where's the evidence? Curr Opin Genet Dev 8:694-700

Smits P, Li P, Mandel J, Zhang Z, Deng JM, Behringer RR, de Crombrugghe B, Lefebvre V (2001) The transcription factors L-Sox5 and Sox6 are essential for cartilage formation. Dev Cell 1:277-90

Sordino P, van der Hoeven F, Duboule D (1995) Hox gene expression in teleost fins and the origin of vertebrate digits. Nature 375:678-81

Spitz F, Gonzalez F, Duboule D (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell 113:405-17

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN (2003) Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21:577-81

Tabin CJ (1992) Why we have (only) five fingers per hand: hox genes and the evolution of paired limbs. Development 116:289-96

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599-607

Tanaka M, Munsterberg A, Anderson WG, Prescott AR, Hazon N, Tickle C (2002) Fin development in a cartilaginous fish and the origin of vertebrate limbs. Nature 416:527-31

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 HJ, Meijlink F, Zeller R (2002) Progression of vertebrate limb development through SHH-mediated counteraction of GLI3. Science 298:827-30

Tellier AL, Amiel J, Delezoide AL, Audollent S, Auge J, Esnault D, Encha-Razavi F, Munnich A, Lyonnet S, Vekemans M, Attie-Bitach T (2000) Expression of the PAX2 gene in human embryos and exclusion in the CHARGE syndrome. Am J Med Genet 93:85-8

Thomas JH (1993) Thinking about genetic redundancy. Trends Genet 9:395-9

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-80

True JR, Carroll SB (2002) Gene co-option in physiological and morphological evolution. Annu Rev Cell Dev Biol 18:53-80

Tyurina OV, Guner B, Popova E, Feng J, Schier AF, Kohtz JD, Karlstrom RO (2005) Zebrafish Gli3 functions as both an activator and a repressor in Hedgehog signaling. Dev Biol 277:537-56

Ueta E, Maekawa M, Morimoto I, Nanba E, Naruse I (2004) Sonic hedgehog expression in Gli3 depressed mouse embryo, Pdn/Pdn. Congenit Anom (Kyoto) 44:27-32

van der Hoeven F, Schimmang T, Vortkamp A, Ruther U (1993) Molecular linkage of the morphogenetic mutation add and the zinc finger gene Gli3. Mamm Genome 4:276-7

Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y (2004) Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A 101:1638-43

Vanderlaan G, Tyurina OV, Karlstrom RO, Chandrasekhar A (2005) Gli function is essential for motor neuron induction in zebrafish. Dev Biol 282:550-70

Vastardis H, Karimbux N, Guthua SW, Seidman JG, Seidman CE (1996) A human MSX1 homeodomain missense mutation causes selective tooth agenesis. Nat Genet 13:417-21

Veltmaat JM, Relaix F, Le LT, Kratochwil K, Sala FG, van Veelen W, Rice R, Spencer-Dene B, Mailleux AA, Rice DP, Thiery JP, Bellusci S (2006) Gli3-mediated somitic Fgf10 expression gradients are required for the induction and patterning of mammary epithelium along the embryonic axes. Development 133:2325-35

Verloes A, David A, Ngo L, Bottani A (1995) Stringent delineation of Pallister-Hall syndrome in two long surviving patients: importance of radiological anomalies of the hands. J Med Genet 32:605-11

Vortkamp A, Gessler M, Grzeschik KH (1991) GLI3 zinc-finger gene interrupted by translocations in Greig syndrome families. Nature 352:539-40

Wang B, Fallon JF, Beachy PA (2000) Hedgehog-regulated processing of Gli3 produces an anterior/posterior repressor gradient in the developing vertebrate limb. Cell 100:423-34

Warburton D, Lee MK (1999) Current concepts on lung development. Curr Opin Pediatr 11:188-92

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L,

Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-62

Werner CA, Dohner H, Joos S, Trumper LH, Baudis M, Barth TF, Ott G, Moller P, Lichter P, Bentz M (1997) High-level DNA amplifications are common genetic aberrations in B-cell neoplasms. Am J Pathol 151:335-42

Westerfield M (2000) The zebrafish book: A guide for the laboratory use of zebrafish (*Danio rerio*) University of Oregon Press, Oregon

Wild A, Kalff-Suske M, Vortkamp A, Bornholdt D, Konig R, Grzeschik KH (1997) Point mutations in human GLI3 cause Greig syndrome. Hum Mol Genet 6:1979-84

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333-41

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJ, Cooke JE, Elgar G (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3:e7

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206-16

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20:1377-419

Yang JT, Liu CZ, Villavicencio EH, Yoon JW, Walterhouse D, Iannaccone PM (1997) Expression of human GLI in mice results in failure to thrive, early death, and patchy Hirschsprung-like gastrointestinal dilatation. Mol Med 3:826-35

Yelon D, Ticho B, Halpern ME, Ruvinsky I, Ho RK, Silver LM, Stainier DY (2000) The bHLH transcription factor hand2 plays parallel roles in zebrafish heart and pectoral fin development. Development 127:2573-82

Zakany J, Kmita M, Duboule D (2004) A dual role for Hox genes in limb anterior-posterior asymmetry. Science 304:1669-72

Zaki PA, Collinson JM, Toraiwa J, Simpson TI, Price DJ, Quinn JC (2006) Penetrance of eye defects in mice heterozygous for mutation of Gli3 is enhanced by heterozygous mutation of Pax6. BMC Dev Biol 6:46

Zhang J, Nei M (1996) Evolution of Antennapedia-class homeobox genes. Genetics 142:295-303

Zhang P, Gu Z, Li WH (2003) Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol 4:R56

Zuniga A, Zeller R (1999) Gli3 (Xt) and formin (ld) participate in the positioning of the polarising region and control of posterior limb-bud identity. Development 126:13-21

PUBLICATIONS

Abbasi AA, Paparidis Z, Malik S, Goode DK, Callaway H, Elgar G, Grzeschik KH (2007) Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers. PLoS ONE 2:e366

Abbasi AA, Grzeschik KH (2007) An insight into the phylogenetic history of HOX linked gene families in vertebrates. BMC Evol Biol 7:239

Paparidis Z, Abbasi AA, Malik S, Goode DK, Callaway H, Elgar G, deGraaff E, Lopez-Rios J, Zeller R, Grzeschik KH (2007) Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. Dev Growth Differ 49:543-53

ACADEMIC TEACHERS

Marburg (Germany)

Grzeschik

Koch

London (UK)

Elgar

Islamabad (Pakistan)

Ahmad

Salman A

Mirza

ACKNOWLEDGEMENTS

All praises to the Almighty Allah, who induced the man with intelligence, knowledge and wisdom.

No words can suffice to express my indebtedness to my supervisor, Prof. Dr. Karl-Heinz Grzeschik for his brilliant versatility, devoted guidance and for being very supportive and dependable, for having immense patience at all stages of this work.

My sincere regards to Prof. Dr. med. Manuela C. Koch, who is extremely supportive of all research activities going on in this department.

Words could not do justice to the help provided by our collaborators:

Zebrafish enhancer / GFP reporter assay:

Prof. Dr. Greg Elgar, Debbie Goode, Heather Callaway
School of Biological and Chemical Sciences
Queen Mary University of London, UK

Chicken enhancer reporter expression analysis

Prof. Dr. Cheryll Tickle, Fiona Bangs
Division of Cell and Developmental Biology
University of Dundee, Scotland,

Mice embryo histological analysis

Dr. Ansgar Schmidt, Ph.D., Sabine Koch
University of Marburg Medical School
Dept. of Pathology, Germany

I am immensely grateful to my friends/colleagues at the Institute of Human Genetics, Philipps University Marburg, for their help encouragement and company, in particular Dr. Zissis Paparidis, Dr. Sajid Perwaiz Malik, Dr. Frank Oeffner, Dr. Beate Achatz, Claudia Moch.

I am thankful to my wife Saneela Amir for helping me in the proofreading and formatting at the final stages of this thesis.

My most fervent thanks are reserved for my beloved wife, parents, brothers, sisters who pray for my success. Without their prayers, support and encouragement it would have been impossible to complete this task/study.

I dedicate this humble effort to my wife Saneela Amir and to my Father and Mother for their endless love.

DECLARATION

Ich erkläre ehrenwörtlich, dass ich die dem Fachbereich Medizin Marburg zur Promotionsprüfung eingereichte Arbeit mit dem Titel "*GLI* genes: cis-acting regulatory elements" im Institut für Humangenetik unter Leitung von Herrn Prof. Dr. K.-H. Grzeschik ohne sonstige Hilfe selbst durchgeführt und bei der Abfassung der Arbeit keine anderen als die in der Dissertation aufgeführten Hilfsmittel benutzt habe. Ich habe bisher an keinem in- oder ausländischen Medizinischen Fachbereich ein Gesuch um Zulassung zur Promotion eingereicht, noch die vorliegende oder eine andere Arbeit als Dissertation vorgelegt.

Teile der vorliegenden Arbeit wurde in folgenden Publikationsorganen veröffentlicht:

Abbasi AA, Paparidis Z, Malik S, Goode DK, Callaway H, Elgar G, Grzeschik KH (2007) Human GLI3 intragenic conserved non-coding sequences are tissue-specific enhancers. PLoS ONE 2:e366

Abbasi AA, Grzeschik KH (2007) An insight into the phylogenetic history of HOX linked gene families in vertebrates. BMC Evol Biol 7:239

Paparidis Z, Abbasi AA, Malik S, Goode DK, Callaway H, Elgar G, deGraaff E, Lopez-Rios J, Zeller R, Grzeschik KH (2007) Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. Dev Growth Differ 49:543-53

Marburg, April 2008

(Amir Ali Abbasi)

CURRICULUM VITAE

Personal Data

Name	Amir Ali Abbasi
Nationality	Pakistan
Address (Pakistan)	Village & P.O Birote, Distt: Abbottabad
Address (Germany)	Institute of Human Genetics Philipps University, Bahnhofstrasse 7 D35037 Marburg

ACADEMIC QUALIFICATIONS

Degree	Subject	Year	Institution
M.Phil	Biochemistry / Molecular Biology	2001-2003	Quaid-i-Azam University, Islamabad Pakistan **Thesis title:** Linkage Studies of Hearing Impairment loci in Pakistani Kindreds.
M.Sc	Biochemistry / Molecular Biology	1999-2001	Quaid-i-Azam University, Islamabad Pakistan
B.Sc	Biological Sciences	1999	Punjab University Lahore, Pakistan
F.Sc	Biological Sciences	1996	Federal Board of Intermediate and Secondary Education Islamabad
SSC	Science	1994	Federal Board of Intermediate and Secondary Education Islamabad

MERITS AND HONOURS

- Currently holding a fellowship (2005-2008) of the Deutscher Akademischer Austausch Dienst (DAAD) and Higher Education Commission (HEC) of Pakistan.
- Merit Certificate for standing first in (M.Sc) Department of Biological Sciences (Q.A.U) (1999-2001): ***Chancellor Gold-Medalist***.
- Merit Certificate for standing first in (M.Sc) Faculty of Natural Sciences (Q.A.U): ***President Gold-Medalist***.
- Obtained First Position in the College in B.Sc and was awarded the ***Merit Certificate***.
- **Graduate Record Examination (GRE, Educational Testing Service, USA)**
  Biochemistry, Cell & Molecular Biology (April 2003).
  Overall %: 87%, Biochemistry: 87%, Cell Biology: 88%, Molecular Biology & Genetics 71%