Long-range transcriptional regulation of vertebrate developmental genes and the evolution of genome architecture

by Pavla Navrátilová

Dissertation for the degree philosophiae doctor (PhD)
at the University of Bergen

2008

Scientific environment

This work has been preformed in

The Sars International Centre for Marine Molecular Biology

Sars Centre

a partner of EMBL

UNI FOB

UNIVERSITETSFORSKNING BERGEN

as part of the PhD program of

Department of Molecular Biology

University of Bergen

UNIVERSITAS BERGENSIS

Acknowledgements

I would like to thank to all people who have helped and inspired me during my doctoral study.

In the first place I especially thank my supervisor Tom Becker for entrusting me with this interesting project. Thanks for his supervision, opinions, as well as giving me extraordinary scientific freedom. Above all and the most needed, he provided me with unflinching motivation, encouragement and support in various ways. I am very grateful to him for allowing me to read the contents of documents, which were not available for me otherwise, sharing correspondence and discussions with other scientists - to let me understand the importance of communication in science and help me to learn the ability of assessment of science quality.

Thanks to Boris Lenhard for taking the responsibility as a co-supervisor. I have also benefited from discussions and collaboration with members of his group at the Computational Biology Unit. Special thanks go to Pär Engström and David Fredman for preparing the zebrafish enhancer database. The collaboration with them was both fruitful and enjoyable.

We were granted very open and, for the project success, crucial cooperation and help from Koichi Kawakami, who kindly provided us with and taught us how to efficiently use the Tol2 transposon system. Another good collaborator was always José Luis Gómez-Skarmeta, who shared experiences and DNA constructs. Many thanks.

I would also like to acknowledge all members and guest researchers of the Sars5 group, especially Øyvind Drivnes, Anja Ragvin and Andrea Weiner who shared their bright thoughts with me, which were very fruitful for shaping up my ideas and also for their friendship and mental support.

Many thanks go in particular to the zebrafish facility people, both the former and present, for their fish-welfare service and understanding our passion for keeping every little glowing fish. Special thanks go to Caterina Sunde for invaluable help with the final phase of my zebrafish screening.

I will never regret working in the Sars Centre, which is an outstanding institute with a special, friendly atmosphere. Thanks to all with whom I have shared this experience in life; the three and half years here were really a turning point in my life. I appreciate financial and technical support provided in this institute. I would like to thank everybody else who was important to the successful realization of my thesis, together with expressing my apology that I could not mention you personally one by one.

Words fail me to express my appreciation to my husband Pavel whose patience, support, love and persistent confidence in me, has taken a load off my shoulder. I owe him for being unselfish and not let his intelligence, passions, and ambitions collide with mine.

Abstract

Despite the recent massive progress in production of vertebrate genome sequence data and large-scale efforts to completely annotate the human genome, we still have scant knowledge of the principles that built genomes in evolution, of genome architecture and its functional organization. This work uses bioinformatics and zebrafish transgenesis to explain a mechanism for the maintenance of long-range conserved synteny across vertebrate genomes and to analyze the arrangement of underlying gene regulation systems. Large mammal-teleost conserved chromosomal segments contain highly conserved non-coding elements (HCNEs), their target genes, as well as phylogenetically and functionally unrelated "bystander" genes. Target genes are developmental and transcriptional regulatory genes with complex, temporally and spatially regulated expression patterns. Bystander genes are not specifically under the control of the regulatory elements that drive the target genes and are usually expressed in different, less complex, patterns. Enhancer detection reporter insertions distal to zebrafish target genes recapitulate their expression patterns even if located inside or beyond bystander genes. We termed these chromosomal segments genomic regulatory blocks (GRBs). To demonstrate, that the regulatory domain of a developmental regulatory gene can extend into and beyond adjacent bystander gene transcriptional units and that these elements indeed regulate target genes, we tested the function of HCNEs around genes encoding transcription factors, *PAX6*, *SOX3* and *SOX11* in both human and zebrafish genomes. Comparing our results with those obtained using mouse, we establish that human elements can be tested reliably in zebrafish. Testing the elements form *SOX11* loci further revealed subfunctionalization after genome duplication and functional turnover as evolutionary processes on the gene regulation. The genome features confirmed by this work were also applied to provide an advance in understanding human mutations causing or predisposing towards genetic diseases. These mutations are frequently associated to the incorrect gene(s) coding region instead of taking the regulatory mutation in distant regulatory elements of another possible causative gene into account. We could demonstrate using our approach that the genes linked to diabetes by genome wide

association study contain in their introns HCNEs regulating different, more distant gene functionally more probably related to the diabetes phenotype.

List of publications

Paper 1: Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS: **Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates.** Genome Res. 2007 May;17(5):545-55.

Paper 2: Navratilova P, Fredman D, Lenhard B, and Becker TS: **Systematic mapping of *cis*-regulatory activity in megabase regions around human transcription factor genes *SOX3* and *PAX6*.** (submitted)

Paper 3: Navratilova P, Fredman D, Lenhard B, Becker TS: **Regulatory divergence of the duplicated chromosomal loci *sox11a/b* by subpartitioning and evolution of enhancers in zebrafish.** (manuscript)

Paper 4: Lenhard B, Ragvin A, Fredman D, Navratilova P, Drivenes Ø, Engström PG, de la Calle Mustienes E, Gómez Skarmeta JL, Tavares MJ, Casares F, Molven A, Njølstad PR, and Becker TS: **Type 2 diabetes susceptibility variants map within long-range regulatory domains of transcription factor genes.** (submitted)

Contents

1. Introduction

1.1 General introduction

Since the completion of high quality genomes of human, mouse, and a growing list of other vertebrate genomes it has become clear that these genomes are very similar both in overall gene content and in the way genes are organized in chromosomal domains. Consequently, our general view of ‘the’ genome is changing. For a long time the majority of DNA was considered to be evolutionary junk. This belief is changing to a view that a genome is a highly sophisticated information space. The number of protein coding genes is nearly the same for human as for a nematode. Each organism must be able to react and adapt to its environment with a finite and, contrary to beliefs held as recent as a decade ago, quite moderate number of genes. These facts hint at the possibility that what was thought to be “junk” may contain information to regulate genes to balance the striking differences in organismal complexities and may also explain the observed high level of adaptability.

From this point of view, it is gene regulation determining why different genes are turned on and off to generate different cells, different tissues and ultimately different organisms. This paradigm underlies not only phenotypic diversity between different species and speciation [1], [2], but altered gene regulation is probably also the cause of intraspecies variation due to quantitative, spatial or temporal effects on gene expression [3].

In spite of this, the phenomenon of gene regulation encompasses diverse mechanisms and so far we do not know about any universal code as is known for protein coding sequences. Moreover, as the protein coding proportion constitutes only a minor part of the genomes (less than 1,5% for the human genome) the non-protein coding part is relatively large. This makes gene regulation an elusive process and the information gained in this field so far is fragmentary. The meaning of “junk” in this context is to be deciphered. A further enigma is the distance at which regulators are able to control gene expression and how genomes are organized. To answer these questions, researchers utilize bioinformatics as well as in vivo model organisms. This thesis utilizes the approach of combining both, and uses zebrafish transgenesis as an alternative to the more traditional mouse model.

This PhD work started in 2005, under very favorable conditions. Zebrafish had already been established as a genetic model organism. I came to a zebrafish laboratory where a large-scale enhancer-detection project was already successfully running. The used transgenic tools, -murine retroviruses, however, were not as efficient or easy, but I was lucky to get my hands on a relatively new transgenesis tool – the Tol2 transposon. Another advantage for the project was the state of the genome assemblies: The human genome was sequenced and close to completion [4], [5], along with many others, and the zebrafish genome project was in its 4th year with initial results published [6]. Whole genome alignments had been recognized as a powerful tool for assessing sequence functionality and an abundance of bioinformatic tools for studying gene regulation was developed. Genome browsers were established that collect various types of genome annotation and experimental data and they became invaluable resources for integration of experimental and bioinformatic knowledge. Progress in sequencing technologies will enable to produce many more animal genome sequences and possibly also more human individual ones [7]. But the emphasis now is being switched from accumulation of data to their interpretation.

1.2 Structure and composition of vertebrate genomes

To date, 22 vertebrate genomes are in full-shotgun or near-complete genomic sequence assemblies, a further 22 are 2x whole-genome shotgun assemblies and 20 BAC (bacterial artificial chromosome)-based sequences of targeted regions in the genome are available [8] not to mention non-vertebrate metazoan species and other eukaryotes, for example [9, 10]. Importantly, sequences are being annotated, which is a demanding, and as yet uncompleted, process. Genome assemblies and their associated annotations are available through genome browsers, such as the ones developed and supported at UCSC [11] (http://genome.ucsc.edu), Ensembl (http://www.ensembl.org) [12] and NCBI (http://www.ncbi.nlm.nih.gov)[13]. The easy access to large-scale data opened scientists a gate towards genomics and allow integrating single pieces of knowledge, understanding genomes structure and drawing principles of genome function. However, we are still short of understanding the general principles and are still locked in a reductionist's position.

1.2.1 Statistics from genome sequencing projects

There are 32 342 genes annotated in the current Ensembl (NCBI36). However, it is broadly suspected that a large fraction of these entries are functionally meaningless RNA transcripts as they show no evidence of evolutionary conservation with mouse or dog and nor with primates either. The zebrafish genome has 17,330 known protein-coding genes annotated (Ensembl zv7); this might conversely be an underestimate, as the transcriptome has not been studied as extensively and alignment to other species reveals many unannotated conserved exons and fragmented genes. This allows the prediction that the number of genes in the human genome will be reduced to a number similar to other vertebrates and comparable to fish [14]. It is, however, probable that human genes are more complex, with much more alternative splicing, alternative promoter usage, non-coding RNA-mediated regulation and trans-splicing, generating a larger number of final protein products [15, 16].

Remarkably, the exons cover only about 1,5% of the human genome (while 45% are repetitive sequences and 53% was called noncoding DNA during the human genome sequencing summary). Non–protein-coding genes (ncRNA), less well defined, such as the ribosomal RNA and tRNA genes involved in protein synthesis or snoRNA for RNA processing, were identified decades ago. More recently, regulatory microRNAs were discovered. A common feature of all known functional RNAs is a secondary structure. During the ENCODE project, analyzed regions were scanned by computational algorithms to predict functional ncRNAs. The result, confirming previous analyses, was 3.7%, which, even though this constitutes a relatively high number, did not explain the surprising level of transcription revealed in ENCODE [17]. More than 75% of the human genome was detected in tiling arrays as primary transcripts and in accordance with this, five- to tenfold more transcription start sites were identified. Interestingly, these frequently overlap with known coding regions or regions previously thought to be transcriptionally silent intergenic regions [18]. Possible explanations could be that computationally unidentifiable RNAs, such as the 17kb long XIST ncRNA gene involved in dosage compensation [19], the novel transcripts encrypt other yet unknown regulatory functions [20] as proposed by Ponjavic et al. [21] who termed them macroRNA. These authors show that some of these RNAs have the same expression patterns as proximal developmental genes and may act as their regulators (J. Ponjavic, personal communication). Some of these new start sites, however, could be annotated as distant uncovered alternative promoters of already known annotated protein coding genes [16]. It is also possible that the RNA products themselves do not have a targeted function, but that their production is connected to other cellular processes such as replication, or to nonspecific influence on gene regulation. An even simpler alternative is to see the abundant transcripts as noise, a result of evolutionarily neutral events that are tolerated by the organism.

These findings together with gene regulation complexity introduced below, challenge the classically viewed term “gene”. However, what stays as a central point in a single definition of a gene is the final functional product (protein) disregarding intermediate products as well regulatory sequences situated in overlapping genomic regions [22].

1.2.2 Evolutionarily conserved sequence elements

The identification of sequences under evolutionary constraint is a powerful approach for inferring locations of functional elements in a genome; mutations of these bases will often be disadvantageous or deleterious to the organism and will be eliminated by purifying selection [23].

Genome alignments have been used since the pre-genome era with both invertebrate and vertebrate sequence data for predicting coding sequence within anonymous stretches of genomic sequence and for inferring the probable function of encoded proteins, but this approach was very early on proposed also for non-coding sequence annotation [24]. Since the first non-coding elements were discovered, many research groups have started to work on conserved (found to be homologous in two or more species) element annotation, but no unified nomenclature or conservation parameters were established. The names and abbreviations were listed and discussed in [25]; in this thesis I use the general term “highly conserved non-coding elements” (HCNEs).

First whole-genome comparisons of human and mouse DNA have provided an invaluable tool for annotation of functional-elements in both genomes [26]. However, ~75 million years (My) of separation from the last common ancestor turns out to be an insufficient time distance to diverge, and especially in regions with high density of HCNEs, human-mouse conservation is too high overall for alignments to usefully single-out specific conserved elements for further study. Human/mouse alignments yield thousands of non-coding HCNEs; as an illustration, at the 90% conservation level there are 265,537 elements of 50 bp or longer. Comparing the human genome to that of non-mammal vertebrates such as fish, which diverged from the human lineage about 450 million years ago [27], is a powerful filter to prioritize sequences that most probably have significant functional activity essential for developmental processes shared by all vertebrates. To compare numbers, there are only 3,127 human/zebrafish 90%/50bp elements [28].

Another fish species representing the vertebrate distal evolutionary extreme is fugu (Takifugu rubipes). Fugu has very compact genome (390 Mb with only less than 15% repetitive sequence – 8x smaller than the human) reducing the search space, but with a similar gene repertoire as human (22 000 annotated protein-coding genes) [12, 29]. Due to the density of functional elements common to vertebrates it is widely used as a reference genome for vertebrate genome annnotation [24] and the human-fugu conserved elements are almost all conserved also in mouse, rat, chicken or zebrafish. To date many of them have been shown in to act as enhancers in reporter assays [30, 31]. Because different regions of vertebrate genomes appear to be diverging at different evolutionary rates, no single type of two-way comparison can be applied with guaranteed success to all genomic loci. As introduced in chapter 1.5, computational biologists have developed numerous tools to search for functional HCNEs through various evolutionary distances.

The position and distribution of HCNEs was found to be non-random as these elements tend to cluster in the proximity (both up-and downstream) of genes involved in developmental processes, for example transcription factors, signaling molecules, receptors and miRNAs [30, 32, 33], often with a density of HCNEs peaking in the vicinity of their target gene regardless of other genes present in the region [28]. Conversely, HCNEs seem to be depleted in immune response loci, around genes involved in oxidoreductase (mitochondrial) activity and structural ribosome components [34].

The investigation of relative distances between single elements showed that they are, unlike distances between genes, significantly conserved [18]. On the contrary, Sanges and colleagues demonstrated that HCNE shuffling takes place during evolution [35]. Unfortunately, these authors neither conclude whether this resulted in any evolutionary change in gene expression pattern nor how frequently this was observed. The largest distance of a regulatory element from the gene it regulates, which was experimentally verified, is around 1 Mb in the case of the human Sonic hedgehog gene, which may still not be the farthest limit [36].

An interesting feature in connection to HCNEs was revealed by genome wide inspection of distribution of ancient transposons [37, 38]. The majority of human transposon-free regions (TFRs) correlate with orthologous TFRs in other vertebrates, despite the fact that most transposons are lineage specific. Most TFRs are not associated with unusual nucleotide composition, but are significantly associated with genes encoding developmental regulators, suggesting that they represent regions of regulatory information needed for precise expression of genes central to early vertebrate development that are unable to tolerate insertions. Alternatively, transposons could interfere with the spreading of H3K27 methylation mentioned in chapter 1.4.4.

Using alignments of 23 mammalian species, the ENCODE consortium presented a set of evolutionarily constrained sequences covering ~4.9% of the human genome. Their annotation analysis illustrated that only about 40% of the moderately constrained sequence represents known protein-coding exons or untranslated regions. To about 20% of these a regulatory function was assigned using ENCODE experimental approaches (i.e. requiring support from at least one line of experimental evidence, including chromatin modifications associated with activation, DNase hypersensitivity, and nucleosome depletion by FAIRE) and the remaining 40% do not overlap with this annotation. Conversely, functional regions (like 60% of promoters or a certain proportion of functional non-coding sequence) lack any conservation at the evolutionary distance level analyzed [18, 39]. For the most part the HCNEs are understood as regulatory protein binding sequences. Sequences bound by one DNA-binding protein are 5-10 nucleotides long, degenerate, and these proteins often act in combinations on a single HCNE, which creates considerable combinatorial potential through a limited number of transcription factors [40]. This suggests that regulatory specificity is dictated not only through the composition, but also by the orientation and spacing between individual binding sites in the HCNE.

1.2.3 Evolutionary processes reflected in HCNEs

Teleosts underwent an additional whole-genome duplication around 330 My ago that coincided with their burst of diversification [41]. This makes fish excellent models in the study of both gene- and non-coding duplication and subsequent events in genome evolution. By alignment of two fish paralogous loci with an outgroup genome, one can comprehensibly demonstrate evolutionary processes such as complementary loss of subfunctions by degenerative mutations in regulatory elements after duplication (duplication–degeneration–complementation model to explain the evolution of duplicated genes proposed by Force [42]), and subfunctionalization by complete loss of one of the regulatory element copies form the proximity of one of the paralogs demonstrated for example by Kleinjan and colleagues [43].

Only a tiny fraction (<0.1%) of mammalian HCNEs are detectably conserved in the genome of the fish, and few, if any, are recognizable within invertebrates such as insects and worms although these groups have their own HCNEs sharing features with those of vertebrates [44] [45]. Ancient genomic duplication events which gave rise to paralogous copies of the genes plus their regulatory sequences is certainly one process of genome-wide regulatory evolution [46]. In support of this, Sandelin and colleagues [32] found five sets of human-mouse-fugu conserved HCNEs to share >75% identity over an alignment length of at least 50 bp. Hovewer, an explosion of non-coding and possibly functional sequences indicates additional processes in evolutionary origin and history of mammalian HCNEs.

About 45% of the human genome is covered by repetitive elements. Bejerano et al. (2006) first reported a clear case of an HCNE family derived from an ancient transposable short interspersed element (SINE) together with experimental data showing one of them to drive tissue specific expression of the nearby *ISL1* gene, which encodes a LIM homeobox transcription factor required for motor neuron differentiation [47]. The same research group later found thousands of mobile elements located near developmental genes undergoing strong purifying selection. They estimated that at least 5.5% of HCNEs have this origin and noted that they occur exclusively in terrestial vertebrates, probably since the coelacanth (400 My), and are being under constraint since 100 My [48]. A similar study by Xie and colleagues [49] lead to the discovery of a further family of HCNEs that was clearly derived from an ancient transposable element. The family includes at least 120 instances in the human genome, most of which are highly conserved in orthologous locations in other mammals. There are also >200 instances in the chicken genome, although most are not in orthologous locations. The family members show high sequence similarity to a zebrafish SINE3 element that is still active [50]. Subsequently, another constrained class of repeats MER121 overlapping with another 900 HCNEs in the human genome was identified [51]. The role of another family of

SINEs in mammalian neuronal specific regulation was experimentally verified by Sasaki et al., [52], who showed this to be the case for *Fgf8* forebrain activity and the brain specific transcription factor gene *Satb2*. These authors suggest that this repeat family is responsible for evolution of mammalian-specific brain development. In concordance with all these examples, comparisons of the genome of the marsupial opossum (Metatheria) with human, mouse, dog and rat (Eutheria) led to the conclusion that at least 20% of eutherian HCNEs are recent inventions that postdate the divergence of Eutheria and Metatheria and 16% of these eutherian-specific HCNEs arose from sequence inserted by transposable elements. Thus, transposons are a strong force in the evolution of mammalian gene regulation. Large HCNE families were distributed by ancient transposable elements, whose sequences may no longer be present or recognizable after decay by neutral evolution, and acquired regulatory roles by exaptation [53]. The process of exaptation could thus be a major driving force underlying the extreme slowdown in the evolutionary clock of a large class of genetic elements.

Adjacent cooperative binding of TFs in HCNE, the TFBSs degeneration and certain degrees of redundancy allow gradual changes, such as TFBSs turnover, which represents a gain or loss of functional transcription factor binding sites [33]. This can happen through simple mutations or small insertions or deletions (indels), and may lead to the rewiring of transcriptional circuits and eventually, but not necessarily (if the compensatory changes occur) to changes in gene expression that can underlie phenotypic changes (reviewed for example in [54, 55]).

The question, widely discussed and crucial for understanding function and evolution of the HCNEs, is which process allows them to be so extremely conserved. HCNEs could be simply mutational cold spots or could be under weak negative selection. These possible explanations were tested on 481 elements in the human genome, at least 200 base pairs long, that are identical to corresponding regions in the mouse and rat genomes, and were termed ultraconserved elements (UCRs) [56]. Despite the high level of conservation, genotyping of different human individuals reveals many single nucleotide polymorphisms (SNPs) within the UCRs. The derived allele frequency spectrum in the UCRs across the human population together with the minimum level of selection required to generate these sequences confirm that the average level of selection on UCRs is lower than that on essential genes. Moreover, for all SNPs tested, individuals homozygous for derived UCR alleles (alleles that differ from the rodent reference genomes) were viable, and healthy [57, 58]. This suggests that mutations in these regions may often have subtle phenotypic consequences that are not easily detected in the laboratory, but when they are numerous, may have a significant cumulative impact on fitness. Conversely, Katzman and colleagues, using the same approach, showed that these regions are under a purifying selection that is three times greater than that acting on nonsynonymous sites in protein coding genes. From this finding, the authors concluded that the UCRs must be functionally essential for development [59]. This disagreement was partially resolved by experiments where four different UCRs were deleted in mice. Over several generations of homozygous mutants no obvious effects on fitness were observed [60]. The UCRs were subsequently shown not to posses any functional features distinct from other HCNEs, as developmental enhancers are equally prevalent in both population of UCRs and other extremely conserved elements. Moreover, the human/mouse UCR set does only partially overlap with human/dog or rat UCRs [61]. Generally, high levels of conservation may reflect only overall size and density of the regulatory function encoded (number of TFBSs), making them easier to detect by current alignment methods.

1.2.4 Conserved synteny

Synteny (from the greek syn- “together” and -tainia “ribbon”) is defined as a set of genes that are on the same chromosome. Conserved synteny then is a situation in which a set of syntenic genes, and other sequence-conserved markers in one species has orthologs that are syntenic (in the same order) on the same chromosome in another species [62]. Identification of syntenic segments between related genomes can facilitate reconstruction of chromosomal evolution and identification of orthologous functional elements. The colinear segments can be further grouped into blocks of large-scale, conserved synteny and with extension of this analysis to more species, a synteny map can be constructed. Palogenomics tries to reconstruct rearrangements and infer the ancestral vertebrate genome [63]. The evolutionary breakpoints observed are non-randomly distributed and cluster at specific “fragile” regions of the genomes [64]. Generally the human genomic areas containing developmental genes and the associated HCNEs represent the largest blocks of synteny, conserved not only in rodents, but all the way to teleosts. This observation reverses the view of evolutionarily breakpoint “hot spots” into the idea of restricted chromosomal breaks due to the requirement to retain regulatory elements in cis- to the target gene [65]. These features, underlying microsynteny appear also in insect genomes although there is no conserved synteny to vertebrates. This establishes that conserved synteny blocks are kept intact as a general principle of development and gene regulation and as a panmetazoan feature [66].

1.3 Genomic regulatory blocks

Delineating boundaries of syntenic blocks from multiple sequence alignments and HCNE density plots help to better understand the boundaries of regulatory blocks surrounding their target genes. The target genes are generally those requiring more complex spatial, quantitative and temporal regulation –often falling into the functional category of developmental regulators. Among these belong especially transcription factors, signaling molecules, microRNA and receptors with their ligands.

In support of the existence of long-range regulatory domains within these genomic regulatory blocks (GRBs), Ahituv et al. analyzed the prevalence and distribution of chromosomal aberrations leading to position effects (disruption of a gene's regulatory environment). They observed a clear bias towards mapping onto and even beyond conserved synteny blocks and also a decrease in gene density which is due to overlap with almost all gene-poor regions in the human genome [67]. The idea of functional links between gene interdigitation and multi-species conservation of synteny blocks was supported by [68] and by a large-scale enhancer detection project in zebrafish, which also helped to recognize target genes.

Different types of GRBs occuring in the genomes are depicted on fig. 1 and further discussed below.

1.3.1 Gene deserts

The distribution of protein coding regions is not random, yielding intergenic regions far longer and more frequent than would be expected by chance alone. Approximately 25% of the human genome consists of gene-free regions greater than 500 kb, termed gene deserts [69]. Many represent cases of GRBs where it is usually straightforward to point at the only target gene regulated by HCNEs scattered throughout the locus.

The first explored gene deserts, 870 kb and 1330 kb in length, bracket the human *DACH1* gene. *DACH1* is an extremely conserved gene involved in the development of brain, limbs, and sensory organs; having features of the target genes highly associated with HCNEs. Nine of the 32 HCNEs form the flanking deserts and *DACH1* introns conserved down to pufferfish were tested in a mouse reporter assay. Seven of these elements were shown to reproducibly drive reporter expression in a distinct set of tissues in transgenic mice, recapitulating several aspects of *DACH1* endogenous expression.

The regulatory activity of HCNEs clustered in gene deserts was shown also for *SOX3* and *SOX11* loci in papers 2 and 3 in this thesis and *SOX10* elements were reported in [70]. Single HCNEs regulating each of these genes exhibit mutually overlapping activities at both the cellular level and in sharing binding sites for transcription factors. This may explain high the level and robustness of expression of these genes during embryonic development.

Nobrega et al. [71] described the deletion of two gene deserts in mice, a 1817 kb region from chromosome 3 and a 983 Kb region from chromosome 19. Together, the two selected regions contain 1,243 human–mouse conserved non-coding elements (more than 100 base pairs (bp), 70% identity), also similar to genome averages, whereas no UCRs or sequences conserved to fish (more than 100 bp, 70% identity) were present**.** The resulting mice were viable with no detectable phenotypic difference and exhibited only minor differences in gene expression compared to the wild type. These results confirm the notion from the UCR deletion mentioned previously [60] and support the idea that the level of redundancy is very high and that these authors might have been unable to detect changes caused by gene desert deletions or that there are compensatory mechanisms for loss of regulatory information. The proposal that gene deserts together with the HCNEs in them have no function in the organism development is contradicted by evidence that some HCNE mutations cause obvious defects, as in the cases described below in chapter 1.7.

1.3.2 Bystander genes in genomic regulatory blocks

Gene deserts represent only a minor subset of genomic regulatory blocks. Large regions spanned by HCNEs often contain genes whose biological functions and expression patterns are unrelated to those of the presumptive target gene. These unrelated genes, which were termed "bystander genes," are independent of the regulatory input of HCNE arrays, but the pressure to maintain HCNE arrays have kept bystander and target genes together for hundreds of millions of years [72], (paper 1). In contrast to the target genes, bystanders can be categorized by function mostly as housekeeping genes and genes with general cellular functions. These genes have not restricted, often weak expression patterns, with supposedly no particular spatial and temporal regulation that would require multiple regulatory elements except for proximal promoters. After the teleost whole genome duplication (and duplication of entire GRBs), duplicated HCNEs were retained with their neighboring target genes, but with lower constraint on the regional integrity due to functional redundancy of the duplicates. This often resulted in loss of some copies of the HCNEs. In situations when the target gene was flanked by bystander genes, one copy of these genes often disappeared by neutral evolution, but left behind the HCNEs formerly residing inside their introns [73]. This supports the idea that these unrelated genes do not receive regulatory input from the GRB in which they reside. Efforts to find differences in promoter sequences of target- and bystander genes did not reveal any– both categories of genes tend to have CpG rich promoters, although in the target gene group it is the majority of them.

The extensively studied *PAX6* locus may serve as an example of this type of GRB. This case is analyzed in paper 3 and its impact on human disease is discussed in chapter 1.7.

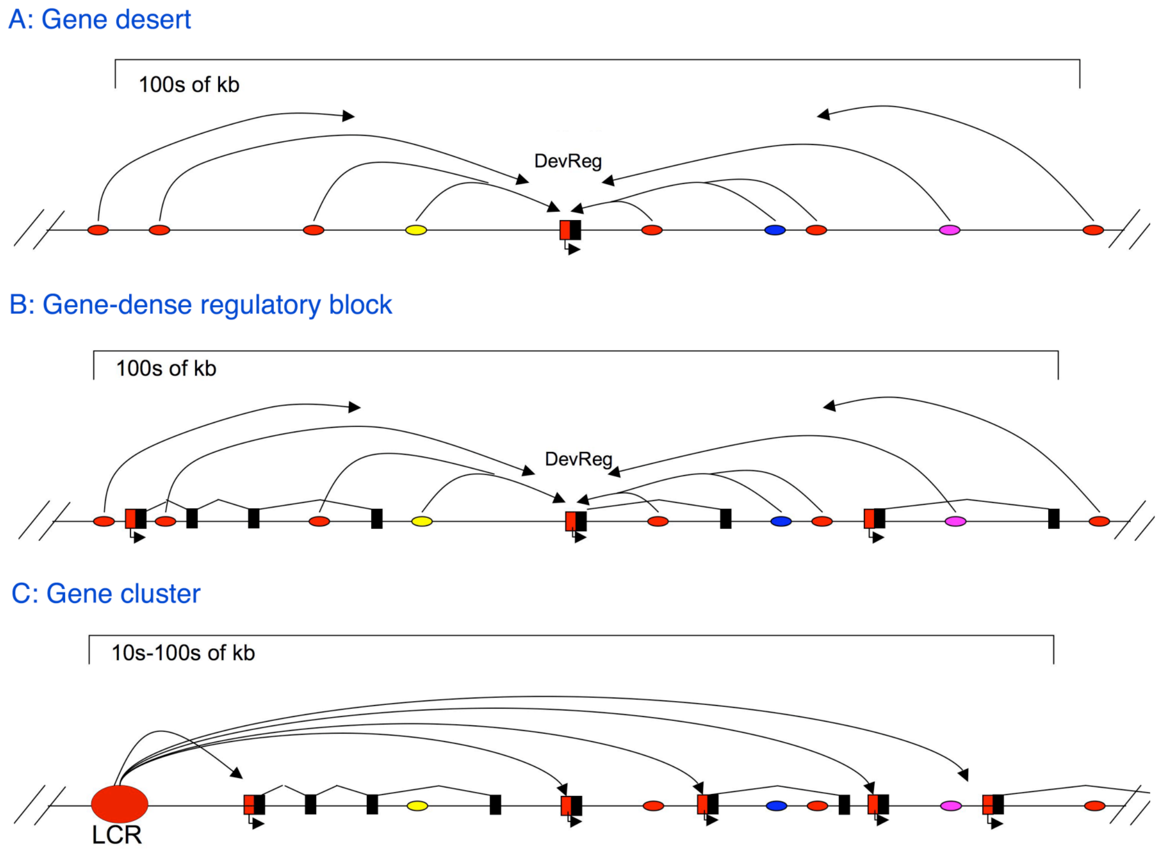
1.3.3 Clusters of co-expressed genes

As a consequence of genome evolution by gene duplication, related genes in vertebrates may be grouped into multigenic complexes. Among these we can find transcription factor gene families like Hox, Six, Fox and Irx, signaling molecules like Fgf and Wnt, but also tissue specific keratins, globins, or large groups of olfactory receptor genes. Genes within these clusters tend to share some expression features and they share regulatory elements. Therefore, their tandem organization is maintained over long evolutionary periods [74]. One specific mechanism keeping gene clusters intact, influencing all involved genes are the so-called global control regions (GCR) or locus control regions (LCR) as in Hox or ß-globin gene clusters. The LCR of the β-globin gene is required for its relocation to the interior of the nucleus. In addition, the LCR mediates the association of the locus with RNA polymerase II (Pol II) transcription factories, which is a common requirement for all genes in the cluster.

Curiously, structurally and functionally unrelated genes sometimes also share regulatory elements. This is the case for *Lnp*, a gene outside of the HoxD cluster, and co-expressed with HoxD genes in the limb. Mutation of *Lnp* was, unlike mutation of Hox genes, demonstrated to have no effect on limb development [75]. It is not known why these genes are kept together when there is no underlying functional constraint.

A genome-wide analysis of the chromosomal distribution of gene expression levels in the human genome indicated the existence of large (~80 genes) regions showing strong clustering of genes of high expression, interspersed with regions where gene expression is low. This was proposed to be an evolutionarily conserved situation where highly expressed domains (called ridges) are gene dense, GC rich, and SINE repeat rich, and the genes have short introns, whereas the weakly expressed domains (called anti-ridges) show the opposite characteristics [76, 77]. Further transcriptome analyses of different tissues uncovered broad (up to 350 kb in size) clusters of spatially co-expressed housekeeping genes [78, 79]. However, such global profiling methods may not reveal partial overlaps of complex expression patterns and are certainly biased towards ubiquitously expressed housekeeping genes. The situations of co-regulated gene clusters were exhaustively reviewed in [80].

Fig. 1: Types of genomic regulatory blocks revealed by HCNE density and conserved synteny evaluation. A: gene-free regulatory block; HCNEs regulate the only target in the block. B: gene-dense block; the target gene receives regulatory inputs from HCNEs, with no respect to the presence of bystander gene. C: clustered target genes share HCNE inputs.



1.4 Mechanisms of gene regulation

Transcription of a protein-coding gene is preceded by multiple events; these include decondensation of the locus, nucleosome remodeling, histone modifications, binding of transcriptional activators and coactivators to regulatory regions and promoters, and the assembly of a pre-initiation complex at the gene promoter, followed by initiation of RNA synthesis, pausing and the transition to productive elongation. Transcriptional regulation is therefore a multistep process that is controlled at the level of chromatin state, recruitment, initiation, pausing, and elongation of RNA polymerase II (RNApolII) [81].

As most of the regulatory information for a cell is encoded in the DNA, regulatory elements including promoters, enhancers, silencers, insulators, locus control regions and other yet unknown elements maintain these processes. The transcriptional enhancers act mostly independently of their orientation and position relative to the promoter [82] which we also found in our reporter studies. All these elements are thought to cluster in so-called chromatin hubs (at least transiently) and eventually make contact with RNApolII molecules, which are distributed as multi-molecular aggregates within the nucleus to form 'factories' for transcription**.** Active genes have been found to loop out of their chromosomal territories upon transcriptional activation. In recent studies, interactions even between different chromosomes have been detected at these factories [83, 84]. The RNApolII stalling was observed exclusively on the promoters of the target genes in *Drosophila* [85], whereas the bystanders, being continually expressed, may not need this repressive step.

1.4.1 Core promoters

The core promoter of a gene includes the transcription start site(s) as well as the region immediately surrounding this site. Various functional DNA motifs, known as core promoter elements, assist in the recruitment, assembly and initiation of the RNA polymerase II (RNA Pol II) transcription machinery. Genome scale methods locating the 5′ boundaries of transcripts or active TSSs indicated that most human and mouse promoters lack the distinct TSS located at one specific genomic position; instead, the typical core promoter architecture consists of an array of closely located TSSs that spread over around 50–100 bp (“broad promoter”). Only small proportions of genes have single TSS. Many hybrids between these two types of promoter also exist; for instance, in some promoters, TSSs are distributed over a large region, but most transcription initiates at one specific nucleotide position. A median broad promoter is 71 bp and its length does not exceed 150 bp [86].

A set of common DNA sequence elements and motifs are associated with core promoters. These patterns have important characteristics that are linked to the expression of the downstream genes. Different elements can co-occur in the same promoter, although certain combinations are more likely than others, and some motifs complement each other. Most of the promoters (50-72%, depending on the definition) are CpG islands associated, mainly those of housekeeping function and ubiquitous genes or brain specific genes and are overrepresented in broad promoters. TATA box motifs are associated with strong tissue-specific promoters, and often co-occur with initiator (Inr)-like sequences at the initiation site. Similar function is exhibited by the DPE element in *Drosophila,* which also occurs together with Inr. Inr as well as the TATA box are the only known core promoter elements that, alone, can recruit the pre-initiation complex (PIC) and initiate transcription. Binding of TBP to the TATA box enforces the PIC to select a TSS in a limited genomic space resulting in a single TSS [86, 87]. An initial attempt to distinguish properties of the target vs. bystander gene promoters bioinformatically failed as they both often contain CpG islands. But after a closer look total length and number of GpG islands overlapped with a GRB target gene and its promoter is larger than by-standers independently of gene length (Altuna Akalin, unpublished results).

1.4.2 Proximal and distal cis-regulation

In compact genomes of simple organisms such as yeast, a gene and its regulatory elements form an uninterrupted genomic segment that forms a "regulatory expression unit." [88]. In higher eukaryotes the proximal promoter is defined as the region immediately upstream (up to a few hundred base pairs) from the core promoter, and typically contains multiple binding sites for activators and repressors which act synergistically. Proximal promoter elements are functionally similar to the distal enhancers, and the distinction between the two classes is somewhat blurred [89].

In order to study gene expression and regulation researchers often used a contiguous stretch of upstream DNA (up to several kb) fused to a reporter gene. However, in comparison to the endogenous expression (by RNA in situ hybridization) the reporter was in some cases misexpressed elsewhere or subpatterns were missing [90, 91]. These data indicate that distal regulatory elements were missing. For this type of studies the more suitable approach is generation of transgenic animals using bacterial artificial chromosomes, which allow to clone hundreds of kb and in many cases is sufficient to recapitulate expression of a gene in transgenic reporter lines [92]. Another option is to use enhancer detection technology, where a reporter transgene with a basal promoter is inserted into the genome [93].

1.4.3 Insulators

Insulators are DNA sequence elements that are thought to prevent inappropriate interactions between adjacent regions of the genome (e.g. promoter and regulatory element) when placed between the two. There are two types of insulator — one that provides a barrier to the spread of heterochromatin and another that is involved in enhancer-blocking activity; some are able to act as both. Several distinct motifs were found to bind insulating proteins in Drosophila, for example *scs* or transposable element *gypsy* [94]. The first vertebrate insulator element discovered is located at the 5' end of the chicken ß-globin locus. A 250 bp DNA fragment, named cHS4, if unmetylated, recruits CCCTC binding factor (CTCF). The same model seems likely to apply to *scs* and *gypsy* insulators: CTCF molecules can interact with each other to form clusters and therefore generate closed loop domains [83]. A genome-wide ChIP-chip analysis described 13,804 CTCF-binding sites in the human genome. These were discovered experimentally in primary human fibroblasts, but CTCF localization seemed to be largely invariant across different cell types. In many cases CTCF consensus motif was shown highly conserved in other vertebrate genomes [95]. Unpublished results by the group of L. Elnitski indicate that only a small proportion of CTCF sites really function as insulators in transfection assays, and that some of them act as enhancers or silencers in this experiment.

Interestingly, a recent study proposed cohesin complexes, which are known to mediate sister-chromatid cohesion in dividing cells, to function as a transcriptional insulator specifically at sites defined by CTCF. Perhaps cohesin is a molecule that structures DNA in a way that causes insulator and boundary effects [96-98].

1.4.4 Current models of transcription and enhancer-promoter interaction

Unresolved questions remain how such elements communicate regulatory effects to their linked genes over large spans of intervening DNA.

Experimental evidence has been provided for a looping model, which postulates that after activators and RNA polymerase II bind to the enhancer, the intervening DNA forms a loop to bring the complex to the promoter [99]. It has also been gradually concluded that the nuclear matrix attachment (to matrix attachment regions – MAR - AT-rich DNA stretches) by cell-specific MAR binding proteins is essential for looping and formation of transcriptionally active chromatin structures [100] which presumably correspond to GRBs. A number of models propose a mechanism of how enhancer-binding proteins and their associated coactivators establish a productive interaction with the cognate promoter: Some propose that enhancers freely 'diffuse' through the nucleoplasm, while the long intervening DNA that does not participate in enhancer function is looped out. This is hard to reconcile with the enhancer-blocking properties of boundary or insulator elements. In the 'tracking' model, the enhancer-bound TF complex actively scans or (tracks via small steps) along the DNA to activate the target promoter [101].

After the discovery of pervasive non-coding transcription, some enhancers were found to be transcribed as well [102] and only by extrapolating EST data onto the interspecies conservation plot, many of them are found to overlap. These findings fit with the facilitated tracking and transcription model [103]. In this model, the enhancer complex tracks through and loops with the intervening DNA to transcribe at low levels short, polyadenylated, intergenic RNAs to ultimately loop toward the distant promoter complex to activate synthesis of mRNA. The interposed insulator traps the enhancer complex in the region 5' of the insulator, thus preventing the enhancer complex from tracking through the insulator and the downstream intervening DNA and activate mRNA synthesis from the distant promoter.

Another plausible explanation for the mechanism by which distant DNA segments get together was recently offered in model of nuclear actin- or myosin-dependent rapid and directed movements of chromosomal loci through the nucleus [104, 105]. In addition, the superhelical tension and chromatin remodeling factors described below appear to influence this process.

Anecdotal evidence exists for interchromosomal interactions between specific genetic loci that are expressed as mutually exclusive alternatives in two different cell types [106]. An example is the olfactory receptor (OR) gene family consisting of >1,200 OR genes scattered over 50 loci in the mouse genome, where only one olfactory receptor is expressed in a particular neuron. One conserved element (H) was demonstrated to associate with promoters of olfactory receptor genes present on different chromosomes and was proposed to function as a universal enhancer for any OR gene in a given neuron. However, deletion of the H element by gene targeting in mice resulted in the loss of expression of members of an olfactory receptor gene cluster according to distance and no effect was observed on the transcription of olfactory receptor genes located on different chromosomes [84]. In *Drosophila* interchromosomal interaction is accepted as a common phenomenon known as transvection, which is the ability of cis-regulatory elements to regulate transcription of the promoter on the homologous chromosome enabled by homologous chromosome pairing [107].

In the global picture, chromatin in the interphase nucleus is compartmentalized into discrete territories and various regulatory proteins are present in specific nuclear bodies and/or are diffusely distributed throughout the nucleoplasm. Chromosomal territories are arranged in a radial fashion whereby gene-rich chromosomes occupy a more central position in the nucleus and gene-poor chromosomes are present closer to the nuclear periphery, partially correlating with expression level. The dynamic translocation of loci upon activation was also observed by several investigators [108].

1.4.5 Epigenetic regulation

The epigenome describes a process whereby potentially heritable changes across the genome are stored as methylation to cytosine bases as well as more than 100 different chemical modifications to the histone proteins that package the genome [109, 110]. By modulating 3D chromatin structure and DNA accessibility, these chemical changes influence how the genome is expressed across developmental stages, tissue types, and disease states.

In vertebrates, DNA methylation occurs almost exclusively in the context of CpG dinucleotides, and most CpGs (except for the CpG islands) in the genome are methylated. Non-CpG methylation (CNG and CNN) has a functional role in plants and might also act in mammals. De-novo methylation occurs mainly during embryonic development, but can also occur in adult somatic cells by aging and carcinogenesis. Methylated cytosines function to promote or preclude recruitment of regulatory proteins. In the former case, the methyl mark can be read through a family of methyl-CpG binding proteins thought to mediate transcriptional repression through interactions with histone deacetylases. Alternatively, the methyl mark can exclude DNA binding proteins from their target sites, as has been shown for CTCF binding at the H19 locus [111, 112]. DNA methylation is dynamically interconnected with histone modification, influencing chromatin packaging.

The CpG islands (CG-rich regions) often overlap with actively transcribed promoters and also a proportion of HCNEs [113]. The CpG islands appear to be unmethylated except for imprinted loci. A small proportion (9,2%) of all CpG islands become methylated during development, and when this happens the associated promoter is stably silent [112]. Other key players in epigenetic regulation of developmental genes are the Polycomb and Trithorax group proteins (PcG) conserved from Drosophila to human. Both Polycomb and trithorax group proteins act to remodel chromatin altering the accessibility of DNA to factors required for gene transcription. Polycomb group genes are involved in chromatin based gene silencing of developmental genes, while Trithorax group genes counteract the silencing effects of chromatin to maintain gene activity during morphogenesis. PcGs form Polycomb Repressive Complexes (PRCs) that function in ES cells to repress genes that are preferentially activated during differentiation and to keep embryonic cells pluripotent. PcG proteins collaborate with a specific set of transcription factors such as *SOX2*, *OCT* and *NANOG* (*POU5F1*), which bind promoters of the target genes. Genome-wide ChIP-chip in undifferentiated pluripotent human embryonic stem cells found the PRC binding domains to coincide with conserved CpG islands in the target gene promoters and with a subset of the HCNEs associated with these genes [114, 115]. The HCNE-clusters spanning developmental genes, boundary conservation among synteny blocks (i.e. the GRBs) and Polycomb binding regions were clearly demonstrated to coincide in Drosophila species [66]. An investigation of individual genes reveals a striking conservation between PcG targets in flies and vertebrates. In the class of transcription factors, 76% of vertebrate PcG targets that have a clear fly homolog and were also identified as fly targets, for example the HOX, PAX, FGF or SOX families [116].

Mechanistically, the PRCs cause trimethylation of Lys27 on histone H3, which facilitates oligomerization, condensation of chromatin structure, and inhibition of chromatin remodeling activity in order to maintain silencing. Recent work suggests that PcG proteins also regulate the nuclear organization of their target genes and that PcG-mediated gene silencing involves noncoding RNAs and the RNAi machinery.

The histone marks are unfortunately too broad an issue, which is beyond the scope of this thesis. A notable point, important for this thesis, is that systematic studies of chromatin modifications have revealed a complex landscape including punctate sites of modified histones at active transcription start sites (H3K4me3), distal regulatory elements and conserved sequences, and broad domains at elongated transcripts (H3K36me3) [112]. Chromatin state can also help to recognize functionally conserved but non-orthologous elements between species as well as novel transcripts, including ncRNA [117].

One should not forget DNA replication, which is a process that has to be coordinated with transcription. The differential replication timing of individual gene categories during S-phase is also connected to epigenetic gene regulation [118]. The unmethylated CpG islands often overlap with replication origins and histone modifications clearly play a major role in this process [18]. We may expect the genomic order of replication sites to be conserved as well with the need of certain sequences accompanied to the process.

1.5 Computational prediction of regulatory elements

The large amount of noncoding DNA compared to the coding sequence, together with possibly different kinds of regulatory information encoded, makes the computational search for regulatory elements difficult. Comparative genomics is the first line approach, representing the most efficient and reliable method.

An alignment constitutes a mapping of one DNA sequence onto another, evolutionarily related, DNA sequence in order to identify regions that have been conserved. There are two basic types of alignment programs, local and global. Global alignments are computed to produce optimal similarity scores over the entire length of the two sequences. Global alignments may be better than local alignments for detecting highly diverged but orthologous subregions in a comparison of two or more long contiguous sequences. Local alignments are computed to produce optimal similarity scores between subregions of the sequences, because the two regions of conserved synteny being compared may have undergone internal insertions, inversions or deletions that preclude an accurate end-to-end alignment. Today, blends of local and global alignment strategies are developed. They are generally based on dynamic programming, handling evolutionary processes of genome divergence like duplications, inversions and indels. A range of currently used algorithms are most recently reviewed by Margulies and Birney [8]. Although it is inherently impossible to know whether the alignment is reflecting real evolution, the existence of different algorithms and comparisons of them with real data are convenient for making a good choice of the tool to use [39].

The method of functional HCNE identification typically performed by multiple global alignment is called phylogenetic footprinting. It discovers regulatory elements by identifying conserved motifs in orthologous regions of multiple species. Computer algorithms designed for this purpose today enable the user to choose optimal evolutionary distance among the sequences under study to make the most accurate predictions. A recent experimental study showing a small proportion of regulatory elements no to be conserved called into question the effectiveness and eligibility of phylogenetic footprintig [119]. Conversely to this study, an ubiased approach by Uchikawa and colleagues [120], who scanned 50kb around the *SOX2* gene showed that all sequences that were positively tested were conserved between chicken and mammals. Of course, the problem might be, as Fisher et al. [121] demonstrate, that not only sequences that are conserved between fish (the organism in which the element was tested) and other vertebrates, but also those conserved only in mammals are able to drive expression in the fish. This observation is confirmed by a highly specific pattern driven by an element conserved only in tetrapods from the *SOX4* GRB, tested in paper 4 in this thesis. Another reason of lack of sequence conservation found in regulatory sequences could be either too large an evolutionary distance used for the alignment or too short an element (a single TFBS), undetectable by existing algorithms. The most supported view therefore is that phylogenetic footprinting is the first-line method and that properly selected non-conserved sequence has no regulatory function.

Phylogenetic shadowing is a variant of phylogenetic footprinting. In contrast to footprinting, phylogenetic shadowing examines sequences of very closely related species and takes into account the phylogenetic relationship of the set of species analyzed. The foundation of this approach is to analyse orthologous sequence from numerous primate species to increase the evolutionary distance of the sequence comparisons. The summation of these primate comparisons robustly identifies regions of increased variation and ‘shadows’ representing conserved segments. This approach enabled the localization of regions of collective variation and complementary regions of conservation, facilitating the identification of coding as well as noncoding functional regions [122].

A new generation of computational tools aimed at predicting tissue-specific regulatory elements, in addition to phylogenetic footprinting and TFBS analysis, also examines gene expression data. Known HCNEs regulating co-expressed genes can be used to identify motifs that are overrepresented in the data set when compared to whole genome. Another way of regulatory element annotation is an unbiased search for overrepresented sequence motifs. Focusing on mammalian-conserved HCNEs, Xie et al. discovered >200 long (12 and 22 nt) motifs enriched in these HCNEs including thousands of CTCF binding sites [123].

In Drosophila, a long history of gene regulation research resulted in a vast amount of experimental data and deep knowledge of principles enables even to “reverse” the bioinformatical analyses to compute and predict gene expression patterns as a function of a set of cis-regulatory sequences, binding-site preferences and expression of participating transcription factors [124].

Several tools to align and visualize whole-genome alignments stand out that are reviewed by Loots [125] and those relevant for this thesis are introduced in the next chapter.

1.5.1 Online bioinformatic tools

It is possible to find a plethora of programs available online. There are several bioinformatics groups that keep developing and improving tools. I will point out some of them, especially those that I used for analyses in this thesis, including the main genome browsers.

The UCSC genome browser offers a set of comparative genomics tracks: 28 vertebrate multiple alignment [126] and chains and nets tracks dealing with larger alignment gaps. The browser also features a human/mouse/rat conserved TFBSs prediction track. The main advantage of these tracks is the possibility of simultaneous visualization of a set of other data such as ChIP-chip results, sequence variation, non-coding RNAs and many others.

VISTA tools visualize custom global alignments generated by MAVID [127] or MLAGAN programs [128, 129]. The VISTA Genome Browser and the ECR Bowser provide precomputed whole-genome pairwise alignments for a number of species together with gene annotation, incorporated into a display. Additionally, the ECR Browser allows the user to change conservation parameters and extract directly the HCNE sequence and search for conserved TFBSs [130].

For the search of transcription factor DNA-binding preferences, these must be modelled as matrices, which are converted into Position Weight Matrices (PWMs or PSSMs) and used for scanning genomic sequences. Combination of TFBSs search with evolutionary conservation is necessary to reduce numbers of false positives produced by simple motif search on a single sequence. The UCSC, VISTA and ECR browsers use the TRANSFAC database (http://www.biobase.de) that contains data on transcription factors with matrices of their experimentally determined binding sites. Another tool for TFBSs shared by two sequences is Consite [131]. This tool uses JASPAR (http://jaspar.genereg.net) a high-quality curated, nonredundant PWMs library.

Ancora (Atlas of Noncoding Conserved Regions in Animals) provides a browser for identifying HCNEs for *Drosophila*, human, mouse and zebrafish as baseline species. It allows the user to identify elements conserved on various evolutionary distances under different conservation parameters. These HCNEs can be viewed in one picture with their density plot, which is extremely useful for estimating the GRB boundaries. The user can visualize in context annotated genes, CpG islands, synteny blocks and other custom tracks [28].

1.6 Biological methods of regulatory element validation

The regulatory elements predicted in silico naturally require confirmation of activity by using biological systems. The bioinformatics, in the program-training process, also await a feedback and large data collections from the biological field to improve the accuracy of future computational tools. Although the set of available methods is growing and progressively becoming high-throughput, the process of functional verification is a bottleneck in extensive non-coding annotation.

Here, I summarize the main in vitro (understood as both biochemical and cell-culture based) and in vivo (using animal models) methods employed for both genome-scale studies and individual elements or limited sets.

1.6.1 Defining DNA-protein interactions in vitro

One of the simplest in vitro methods to demonstrate binding of proteins to DNA is the electrophoretic mobility shift assay (EMSA). At most EMSA can tell us that a transcription factor is able to bind to a sequence in the test tube but does not necessarily reflect the *in vivo* situation. Nevertheless EMSA can serve as a supplemental experiment to more specific results [132]. The binding preferences of transcription factors are derived also from another in vitro technology termed SELEX (Systematic Evolution of Ligands by Exponential Enrichment), whose results became the basis for deducing most PWMs [133].

A method to locate active functional elements in cultured cells is through the identification of DNA regions in which nucleosomes are temporarily displaced and thus hypersensitive (HS) to DNase I cleavage. Until recently only individual HS sites could be detected using traditional Southern blot assays [134]. Two modifications of the method allow identifying HS sites genome-wide: they start with the DNase I treatment of chromatin, followed by the attachment of a biotinylated linker to the DNase I-digested ends. The linker is used to extract short adjacent DNA fragments that can be identified by either next generation sequencing (DNase-seq) [135] or labeling and hybridization to tiled microarrays (DNase-chip) [136].

Crosslinking proteins and DNA with formaldehyde directly in the cells and subsequent immunoprecipitation using an antibody against DNA binding proteins (Chromatin immunoprecipitation - ChIP) has recently become one of the most productive methods. ChIP permits to study not only all types of cis-regulatory elements but also histone modifications such as acetylation and methylation. Especially in combination with microarrays and sequencing ChIP provides DNA-protein binding information on a whole genome scale with high specificity [137, 138]. Similar results but without the, sometimes difficult, use of antibodies is provided by the DamI technique [139].

Several experimental approaches for analysis of the mechanisms of communication over a distance between DNA regions positioned on the same molecule and, in particular, for analysis of the mechanism of enhancer-promoter communication were developed recently. Chromosome conformation capture (3C) is a relatively new promising method for linking TSSs and regulatory elements to their corresponding transcript. This technique detects the physical interactions between chromosomal regions that are involved in common regulatory mechanisms. As for the ChIP and DNaseI techniques, microarray, sequencing and other adaptations of 3C allow scanning whole genomes for regions physically close to an analyzed locus (reviewed in [140]).

An interesting technique to tag and recover chromatin in the immediate vicinity of an actively transcribed gene is RNA TRAP (tagging and recovery of associated proteins) which is a modified RNA fluorescence in situ hybridization method to study interactions between distal elements regulating transcription [141]. In the near future, we will undoubtedly be able to witness new insights into 3D aspects of gene regulation.

1.6.2 In vivo testing of regulatory elements

Gene regulation was initially studied mainly in bacteria and many basic principles from eukaryotes have been derived with the help of yeast genetics, but with an interest in tissue-specific gene expression I will focus here on metazoan and especially vertebrate model organisms.

Transcriptional regulation can be studied by enhancer detection (ED) technology based on random integrations of a minimal promoter and reporter gene into the genome. In ED, the minimal promoter is activated by cis-regulatory elements present within reach of regulatory elements and reporter gene reflects, at least partially, the expression pattern of the target gene. Once an expression pattern is recovered, the insertion can be located by isolating sequence flanking the insertion and by mapping it to the genome sequence. Another way of exploring regulatory landscape is through reporter expression, driven by tested regulatory sequence either in transient assays or after integration of the test construct into the genome.

Taken the, albeit distant, common ancestry and obviously shared principles of gene regulation, research focused on vertebrates can certainly take lessons from the techniques used, and results achieved, in studies in the fly. Genetic methods like P-element or enhancer detection were efficiently used even before the sequencing of the Drosophila melanogaster genome to identify regulatory elements and understand their function [9, 142-144]. Today, its sequenced genome together with those of other Drosophila species and insects enable comparative genomics and the regulatory annotation coverage is remarkable [9, 145].

The ascidian chordate Ciona, with its two species sequenced, ensures that the non-coding search space is reduced comparing to vertebrates. The developmental regulatory networks of genes largely homologous to the vertebrate set are well studied and, finally, electroporation of reporter constructs into developing embryos makes in vivo expression analyses of enhancers very fast [109, 146-148]. Recently ED based on the Minos transposon was employed to enable enhancer detection [149].

However, Ciona has a very simple body plan and lacks many vertebrate structures, which is limiting the desired extrapolation to higher vertebrates.

Chicken, the first sequenced non-mammalian tetrapod, has an optimal evolutionary distance from human, and thus provides high specificity in detecting functional elements, both non-coding and coding and has been successfully used for testing them [120, 150]. However, the technique of in-ovo electroporation of DNA constructs results in highly mosaic expression of the reporter and thus decreases the value of this model organism. This shortcoming may be overcome by the use of transposons in the near future [151].

The successful ED projects in Drosophila were early on followed by attempts in mouse [152] which, however, due to demanding ES cells technology ruled out genome scale projects [153]. Already twenty years ago, Kothary et al. [154] created mouse heat shock 68 promoter ß-galactosidase (LacZ) reporter gene transgenes to study heat shock inducing factors. They aimed at enhancer detection and the method of transient transgenesis with this construct was later transformed into a tool for testing tissue specific regulatory elements. A DNA construct containing the element with a promoter and lacZ is linearized and injected into fertilized mouse oocytes, which are then reimplanted into pseudopregnant females. Embryos collected at a certain stage (usually between E10.5 and 14.5) are stained for LacZ reporter gene activity. Pennacchio et al. [155] published a set of 167 human HCNEs that are conserved in human-pufferfish, fugu, or ultraconserved in human-mouse-rat and reported that 45% of these sequences functioned reproducibly as tissue-specific enhancers of gene expression at embryonic day 11.5. To date, the dataset obtained by this group includes 806 tested HCNEs from which 41% are positive. This low frequency reflects the main disadvantage of the single time point of investigation and the likely possibility that a fraction of the negatives may be enhancers active either earlier or later in development (in zebrafish it is around 70% positives). Mouse also offers the possibility to use large genomic constructs such as BACs. Targeted knockouts, routinely preformed exclusively in mouse, allow deletion of regulatory elements, and loxP or FRT recombination would facilitate studying putative interactions between regulatory elements. All the techniques performed in mouse are usually quite costly and technically demanding.

Using transparent animals developing outside the mother allows continuous observation and circumvents the single time point problem and sacrifice of the mother. The clawed frog Xenopus tropicalis has some of these advantages although it is not totally transparent. Established approaches of transgenesis in Xenopus include restriction enzyme-mediated integration (REMI) [156] and the I-SceI meganuclease system [157]. The advantage is that, like in mice, the injected embryo (F0) provides the result and also that large constructs can be tested in this way. A number of laboratories have tried transposon based systems, such as Sleeping Beauty, Tol2 [158], and phiC31 integrase [159], to create stable transgenic Xenopus, but results from these recent studies are not yet available.

1.6.3 Fish as an animal model for exploring the function of regulatory elements

The use of zebrafish overcomes many problems of the above model organisms. High numbers of this small teleost species can be reared in a relatively small space and at a reasonable cost. This model has a short generation time (2 - 3 months from fertilization) and crossing and raising routines are simple. The total transparency of developing embryos permits the use of fluorescent proteins as a reporter gene not requiring terminal fixation for visualization. This allows to follow the spatiotemporal specificity of cis-regulatory information of many GRBs detected in a large- scale ED project performed by Ellingsen and colleagues [25]. In that project, the number of ED transgenic lines came up to 1000 and provided invaluable tools for research on gene expression and -regulation. Although this technique in itself would not allow identification of individual regulatory sequences, the combination with phylogenetic footprinting makes it a powerful approach for defining cis-regulatory activity in a precise genomic location and permit measuring the extent of the regulatory domain of a given enhancer for a given target gene. The HCNEs from this region can be subsequently tested as performed in this thesis.

Zebrafish is also very accessible through various efficient transgenic tools. From all tools functioning for the species introduced above, nearly all were successfully applied to zebrafish. Among the most efficient was an engineered murine leukemia retrovirus (MLV) (described in [160]), which was the means to insert the ED constructs in the project by Ellingsen et al. The high efficiency of viral integration is however dearly bought by the procedures necessary to produce high titer retroviruses for injection. For testing hundreds of HCNEs, more suitable transgenesis methods were therefore needed. The Tol2 transposon was identified in medaka fish [161] and employed by Kawakami in zebrafish and was later also demonstrated to work in Xenopus, chicken, mouse and human cells [162]. Transposase-recognized sequences of the Tol2-based vectors contain essential terminal inverted repeats and subterminal sequences. DNA inserts of fairly large sizes (as large as 11 kb) can be cloned between these sequences without reducing transpositional activity. Cloning of the HCNE can be performed effortlessly by using the Gateway® system upstream of a minimal promoter. Our laboratory uses the zebrafish gata2 promoter sequence, identical with the one used in the ED project. Some groups use either the endogenous promoter of the target gene [74], or other heterologous promoters, but as I show in the supplemental results, the enhancers seem to be promiscuously activating different promoters, suggesting that any promoter of sufficient strength can be used. Transposase mRNA preparation is simple and microinjection of DNA and mRNA is straightforward and can be mastered in just a few days of practice. Therefore, large numbers of transgenic constructs can be injected for high-throughput analysis of regulatory elements. The F1 screening for GFP positive transgenic lines yields 20-90% success, and given the variability due to position effects, 3 to 8 independent expression patterns should be documented to confirm element specificity. Attempts to test HCNEs in transient expression (documenting F0 with or even without Tol2) result in mosaic non-specific expression, inconsistency and low resolution of the data and are suitable only for rough estimates of element activity [30].

To introduce large constructs like BAC sequences into the zebrafish genome, one has to go back to the traditional DNA microinjection with a germline transmission rate of 2% or less [163]. The benefit of using large constructs would be to gain insight into possible enhancer interactions and also for the annotation of silencer activity as well as reduction of position effects.

Medaka (Oryzias latipes), a small freshwater fish, has advantages for genetic and developmental studies similar to zebrafish such as a small body size (3 cm in length), external fertilization and development, transparency of eggs, short generation time, facility of crossing and breeding, and established methods of creating transgenics. The Tol2 system is naturally supressed in its host organism and cannot be efficiently used. An advantage is the more compact genome compared to the zebrafish, which means closer positioned cis-regulatory elements allowing their analysis in combinations without reducing the spacing DNA. A report by I. Conte and P. Bovolenta [164] provides the first example of the precise regulatory code necessary for the expression of the *Six3* gene. These authors analyzed a cluster of conserved noncoding blocks contained within the first 4.5 kb upstream the gene. By testing a series of constructs carrying different combinations of HCNEs combined with an EGFP construct, these authors demonstrated the functional interplay of HCNEs. Such analysis could identify not only enhancers, but also silencers, and silencer blocking activities that are combined to control the distribution of *Six3*.

1.7 Non-coding mutations and human disease

For many genetic diseases, mutations have been characterized in the coding region of the causative gene. Less recognized than the integrity of a protein coding sequence is the fact that proper function is also dependent on the spatial, temporal and quantitative correctness of gene expression and that, if disrupted, such lesions could underlie many human diseases. Regulatory mutations contributing to human disease have been reported. Most of them affect promoter regions whose precise location is known for many human genes.

Another situation is position-effect human disease, i.e. disease associated with chromosome rearrangements that change a gene's position, but do not change the gene’s sequence. These can be caused either by position effect variegation – bringing the gene into different chromatin environment – or by disrupting the regulatory mechanism that ensure its proper expression. Contribution of distant acting mutations to human disease has so far not been explored on a large scale, but considering the large regulatory domains of developmental genes, it might be a common phenomenon. One of the few known examples is the group of limb deformations mapped to the Sonic hedgehog (*SHH*) GRB. Besides its function in ventral midline and central nervous system, *SHH* acts as a morphogen in defining the anterior-posterior axis in the developing limb and the long-distance enhancer ZRS is crucial for its proper dosage in limbs. This element is conserved in vertebrates with limbs or limb-like structures such as wings or fins, but is absent in limbless species such as snakes. As it is located at the extreme distance of one megabase from the gene it regulates, residing in the intron of a neighboring gene LMBR1 (Limb region 1 homolog), the initial study of this locus made the –erroneous- conclusion that LMBR is the causative gene [165]. Translocation, insertion, duplication and even point mutations affecting this element are sufficient to cause limb deformities such as polydactyly (OMIM 174500), syndactyly or acheiropodia both in human and mouse [36, 166]. Detailed studies by mouse transgenic assays revealed the real mechanism. It was shown that single nucleotide substitutions operate as gain-of-function mutations that activate SHH expression at an ectopic embryonic site; and that the sequence context of the mutation is responsible for a variable regulatory output [167]. Elimination of entire 1167 bp conserved intronic region results in complete loss of the autopod truncations in mice, suggesting that the type of mutation determines the severity of phenotype [168]. Interestingly, a recent study associated regenerative failure in the Xenopus adult limb with methylation status of this enhancer region of Shh. Thus methylation might be another mechanism of decreasing the regulatory activity of enhancers [169]. Translocations up to 300 kb from SHH which disrupt communication with enhancers that control SHH expression in the forebrain cause holoprosencephaly (OMIM 142945), a structural malformation of the brain, demonstrating how complex patterns are partitioned to multiple regulatory elements [170].

Another example to show further principles connected to regulatory disease is aniridia, a congenital eye malformation (OMIM 106210) caused by haploinsufficiency of the *PAX6* gene. The haploinsufficiency results from a strict gene-dosage requirement of *PAX6* when this is lowered by protein coding mutations or deletions of one copy, or increased in mice carrying multiple copies of the gene [171, 172]. In the mouse, heterozygosity for mutant *Pax6* is the cause of the small eye (Sey) phenotype.

In some aniridia patients, however, the *PAX6* gene remained intact, but 3' located chromosomal breakpoints were characterized [173, 174]. The 3' region of the *PAX6* contains a dense array of unrelated genes; the neighboring *ELP4* is a ubiquitously expressed gene encoding a protein that associates with histone acetyltransferases and RNA polymerase II to aid transcriptional elongation, and most of the breakpoints map within its introns. A further downstream gene, *IMMP1L,* has peptidase activity and is localized to the mitochondria. Both *ELP4* and *IMMPL1* are unlikely to be related to aniridia. Instead, through a yeast artificial chromosome (YAC) rescue of the Sey mutant mice, essential HCNEs within *ELP4* introns were identified [175]. Their *PAX6* specific enhancer function was also demonstrated by a reporter assay in mouse and gave evidence that removal of these elements causes *PAX6* downregulation. Larger deletions encompassing not only *PAX6*, but also another developmental regulatory gene, *WT1* (Wilms' tumor gene 1) cause a contiguous syndrome termed 'WAGR' (OMIM 194072, named for Wilms' tumor, aniridia, genital or urinary tract abnormalities, and mental retardation) [176]. The *PAX6* and *WT1* GRBs are kept in conserved synteny together with *RCN1* (encoding a Ca2+-binding protein of the endoplasmic reticulum), a bystander gene in between them [177]. HCNEs in both downstream *PAX6* bystander genes (in *ELP4* and *IMMP1L*) and towards *WT1* were tested in zebrafish as shown in paper 2 of this thesis.

The study of a pedigree with familial hypoparathyroidism (OMIM 307700) in an X-linked recessive form revealed a 25 kb deletion around 67 kb downstream of the *SOX3* gene [178]. Taken their position and expression pattern, the regulatory elements analyzed in paper 2 of this thesis are highly probable culprits of this disease.

The cases mentioned above exemplify the loss of one or more enhancers, but other cases could also result from loss of a repressor or gain of inappropriate regulatory elements. The regulatory disease can display in both homozygous (recessive) and heterozyous (semidominant or dominant) states. An exhaustive summary of these and many other identified regulatory diseases is provided by Kleinjan et al. in a recent review [179].

1.7.1 Genome-wide association studies and GRBs

The DNA sequence of any two people is 99.9 percent identical. The variations, however, may greatly affect an individual's disease risk. Sites in the DNA sequence where individuals differ at a single DNA base are termed single nucleotide polymorphisms (SNPs). Sets of nearby SNPs on the same chromosome are inherited in blocks between two recombination sites. This pattern of SNPs on a block is a haplotypes, which may contain a large number of SNPs, but a few SNPs are enough to uniquely identify a haplotype block. Most chromosome regions have only a few common haplotypes (meaning each with a frequency of at least 5%), which account for most of the variation from person to person in a population. The international project called HapMap scans human populations to identify a map of these haplotype blocks and the specific SNPs that identify the haplotypes are called tag SNPs; currently 3.1 million tag SNPs are publicly available [180]**.** Together with microarray high-throughput scanning of genomes this collection has already allowed genome-wide linkage disequilibrium mapping of common varaints in the human population that confer risk to diseases such as diabetes, schizophrenia, cancer and others, especially previously intractable multifactorial disorders; e.g. [181-183].

Efforts to understand the molecular mechanisms involved in development of type 2 diabetes led to the detection of SNPs highlighting susceptibility loci. Some of them clearly indicate the SNP-linked gene’s functional relationship to the disease, such as a polymorphism in IGF2BP2 (insulin-like growth factor 2 mRNA binding) gene. Sometimes, there are two possible culprits as in the locus of the genes *HHEX*/*IDE* (homeobox, hematopoietically expressed and insulysin) [184, 185]. In other cases, the gene closely linked to the tag SNP has a role in diabetes pathogenesis. A SNP located in the intron 5 of the *CDKAL1* (cyclin-dependent kinase 5 regulatory subunit associated protein 1-like 1) has no clear role in pancreatic function [186]. One report proposes a function of *CDKAL1* in regulation of insulin secretion [187], but in light of the evolution of genomic architecture, we also have to view the *CDKAL1* gene as a bystander gene of the neighboring *SOX4* gene, which encodes a transcription factor involved in pancreatic ß-cell development [188]. This situation is analyzed in paper 4 of this thesis along with the case of the *FTO* gene, highlighting a haplotype block highly associated with obesity [189]. *FTO,* like *CDKAL1,* contains numerous noncoding elements in its large introns and is located adjacent to a gene desert next to *IRX3*, a transcriptional repressor expressed in multiple tissues, including hypothalamus [190].

Remarkable help in search for disease loci is also brought about by the dog genome sequencing project [3]. Many modern dog breeds show a high prevalence of specific diseases common in humans. The high prevalence of specific diseases within certain breeds suggests that a limited number of loci underlie each disease, making their genetic dissection potentially more tractable in dogs than in humans. A team from the Broad Institute analyzing the sequenced dog genome generated a dense map of >2.5 million SNPs across 12 breeds to facilitate genome-wide association studies to identify functional sequences responsible for diseases and traits which may be similar or identical in humans [191].

2. Aims of the study

During the collection of hundreds of enhancer detection transgenic zebrafish lines, the question emerged what is the mechanism selecting which of the genes near the transgene integration will be reflected in the reporter expression pattern. The analyses of expression patterns of these lines together with a detailed bioinformatic analysis of the genomic context of reporter insertion site were used to reveal general principles of putative gene regulatory information organization. Results from this approach have initiated the idea of genomic regulatory blocks. However the hypothesis of interdigitation of target gene regulatory elements with bystander genes needed to be proven. When the project was started, only a handful of enhancers were tested in the zebrafish and most of them only by the transient trangenesis. To test the putative regulatory elements of the target genes, I applied the novel transposon-mediated transgenesis to zebrafish to pursue analysis of detailed spatial and temporal specificities of these elements in stable transgenics. More general aims were to answer some of the questions regarding the evolution of gene regulation, mechanisms in promoter-enhancer communication and the overall organization of regulatory information in genomic regulatory blocks. Work to which this thesis also contributed was aimed at resolving ambiguities of genome wide association studies by proposing and demonstrating the extent of gene regulatory blocks of distal genes overlapping with detected haplotype regions.

3. Summary of results

3.1 Enhancer detection screen in zebrafish is a tool for visualisation of genomic regulatory blocks

(Paper 1)

Ellingsen and coworkers [93] devised a retrovirus-based insertional system using yellow fluorescent protein (YFP) under the control of the zebrafish gata2 promoter that allows visualization of the cis-regulatory information in the region where the insertion has occured. In this large-scale enhancer detection in zebrafish, hundreds of insertions were isolated. Here we analyzed expression patterns of subset of these transgenic lines together with mapping activated reporter construct integrations. We revealed significant overrepresentation of developmental regulatory genes among the loci detected, sometimes detected by multiple insertions scattered within the genomic area (e.g. in genomic area of the *id1* gene, nine insertions within ~50 kb were mapped). The target genes were often found distant from the insertion site, leaving even closer located or host genes undetected and not reflected in the expression pattern. This indicated that the unique cis-regulatory content of a large region is devoted only to a single 'target' gene. With the help of comparative genomics, we further demonstrated that the target genes are located in the largest vertebrate blocks of conserved synteny encompassing multiple unrelated genes. Further, the conserved syntenic blocks are rich in highly conserved non-coding elements (HCNEs), exhibiting significantly higher density in the vicinity of the target genes detected in the enhancer trap screen. The conserved non-coding sequences are known to act as specific developmental enhancers and can be found located over large regions, and often distributed in the introns of several adjacent genes to the target gene. In this context, we proposed the term 'bystander' genes as they appear to be unaffected by the regulatory content of the block and usually are functionally and by expression patterns unrelated to the target gene. The regions of conserved synteny containing target genes, bystander genes, and regulatory elements were termed genomic regulatory blocks (GRBs). The concept of this type of genomic arrangement was further supported by the analysis of the zebrafish duplicated genomic regions. Unlike single-copy GRBs, after the duplication, bystander genes were often lost from originally intact synteny by evolutionary decay, but notably leaving behind the HCNEs formerly located within their introns. The GRB organization was demonstrated not only on protein-coding genes, but also microRNA genes, which under this criterion can be included into the category of developmental regulators.

3.2 Analysis of the GRBs organization by functional testing HCNEs

(Paper 2)

To prove that regulatory elements can be located distantly to the target gene and reside in the introns of bystander genes within a GRB, I used zebrafish reporter transgenic assays and devised a high-throughput test for regulatory elements. I utilized the efficient Tol2 transposon to deliver HCNE sequences joined to a minimal promoter and a GFP reporter sequence into the zebrafish genome, which allowed the production of stable transgenic lines. For each element tested, I collected and analyzed multiple lines confirming the expression pattern driven by the tested sequence. Zebrafish enhancer detection lines characterized previously together with RNA in situ hybridizations of genes present within the GRB enabled the comparison of expression to that of each element tested and allowed to assign HCNEs to their target gene.

As a first approach, a multiple alignment-based determination of minimal conserved synteny and HCNE density plots were generated for each region. We showed on the *sox3* locus that in gene-poor regions (so called gene deserts) that multiple HCNEs regulating one central target gene can be easily identified and that the resulting patterns exhibit a high level of overlap.

Some of the *sox3* enhancers are situated more distantly to their target gene than to a neighbouring bystander gene. I have chosen the element SOX3_hs8A, which is twice far away from *sox3* than from the unrelated gene *atp11c*, to test whether the promoter of *atp11c* will be activated by a *sox3* enhancer, and whether the promoter sequence itself provides a mechanism for prevention of inappropriate activation. The expression pattern obtained was unchanged in specificity, but much lower in the intensity. As I did not find evidence for any functional insulator in the genomic region between SOX3_hs8A and *ATP11C* (the only CTCF binding site identified previously did not block specific enhancer-promoter interaction), I could not conclude that the bystander gene was protected from activation by *sox3* enhancers. After testing this element in a reporter construct containing the human *SOX3* promoter, the SOX3_hs8A enhancer was combined with four other promoters. In stable transgenes, all enhancer-promoter combinations resulted in reproducible expression patterns nearly identical to the combination with the natural human *SOX3* promoter. In conclusion, the human enhancer SOX3_hs8A activated a range of heterologous minimal promoters in a *SOX3* specific pattern.

The genomic regions bearing *pax6a* and *pax6b* in zebrafish represent gene-richer GRBs arisen by whole genome duplication in teleost lineage from a single ancient *PAX6*. In human- and other vertebrate genomes that diverged before the whole genome duplication event, the downstream region contains bystander genes with large introns (*ELP4, IMMP1L* and others) containing many HCNEs. In zebrafish, one copy of each bystander gene was lost by neutral evolution and HCNEs present in human genome could be found in the former intronic areas. Testing some of these elements in our assay verified them to act as elements with *PAX6* regulatory function conserved in evolution. In the upstream region, *PAX6* neighbours another target gene, *WT1*, as could be determined by HCNE density plot and developmental function of that gene. Between those two genes, HCNEs tested resulted in both *PAX6* and *WT1* targeting enhancers.

3.3 Evolutionary processes at the level of cis-regulation

(Paper 3)

A further GRB that was analyzed for regulatory content was that of human *SOX11* and its 2 orthologous regions in the zebrafish genome that were retained after whole-genome duplication. Sequence alignments revealed differential loss of HCNEs between the *sox11a* and *sox11b* GRBs, suggesting that degeneration and complementation processes were responsible for subfunctionalization of the two regions. Eight out of nine human/zebrafish HCNEs were found kept in the *sox11a* GRB and only three out of nine in the *sox11b* region. Two of the zebrafish elements were preserved in both paralogs. Human HCNEs showed reproducible *SOX11* specific activity in six cases out of nine, one produced specific, but less consistent patterns and other two were inconclusive. The orthologous elements in the zebrafish *sox11a* region resulted in a lower proportion of highly specific elements resulting in consistent patterns – we found only two among the six that were tested, three were less consistent but specific for *sox11a* gene expression. All three elements in the *sox11b* region were highly specific enhancers whose patterns were consistent with these elements being *sox11b* enhancers.

Testing the two HCNEs duplicated in zebrafish and comparing human vs. zebrafish enhancer-driven expression offered the possibility to test whether extant paralogous sequences functionally diverged following the whole genome duplication and also to determine functional divergence to the human enhancers. A pattern divergence was apparent for both situations. Finally, we discussed these results with sequences analysis and concluded that lower sequence identity between homologous sequences shows a trend towards less consistent patterns obtained in our transgenic assay and that the divergence between paralogous elements is even larger than between human/zebrafish orthologs despite the extreme evolutionary distance between the two organisms. This may be due to higher substitution rates in the duplicated teleost genome after the relaxation of functional constraints.

3.4 Importance of GRB for human disease-genome wide association studies

(Paper 4)

We demonstrated genomic regulatory blocks as a type of genomic arrangement that is common around genes encoding developmental transcriptional regulators. This fundamental feature of vertebrate genomes is also essential for the understanding of human disease mutations in non-coding sequence.

We reviewed genome wide association studies data for human type 2 diabetes and obesity (T2D/O) risk, which usually link SNPs overrepresented in this population to the nearest gene. Our analysis of genomic areas containing T2D/O risk haplotype variants uncovered them to fall into GRBs containing high densities of HCNEs. We could associate the disease-linked SNPs in those regions to new possible candidate genes, which play role in Wnt signalling, and in pancreatic or hypothalamic development. We found HCNEs within the disease associated genomic regulatory block and tested some within the disease associated haplotype region in the zebrafish reporter assay. The expression patterns driven by the human elements supported our hypothesis that the tested HCNEs do not regulate the nearest (by the GWA studies proposed) genes, but the GRB target. While the function and effect on risk of T2D of *CDKAL1* (CDK5 regulatory subunit associated protein 1-like 1) is unknown [186], *SOX4* was proven not only as vertebrate pancreas developmental gene, but also one of the determining factors for adult insulin secretion [188, 192]. Even though conserved only in mammals, the human element located in close proximity to the SNP directed expression in reproducible *sox4* pattern.

This generally implies that not only coding mutations near the disease associated SNP can be causative, but possible distant regulatory mutations of the linked GRB target gene should be considered as a disease mechanism behind common genetic disease.

3.5 SOX3 enhancer–deletion study

(Supplementary results)

The enhancer SOX3_hs8A, which had a highly reproducible and specific activity (in epibranchial placodes, olfactory pits, telencephalon, hindbrain, inner ear and spinal cord), was used for fine-scale analysis of functional motif composition. I created a number of deleted variants of the enhancer, covering the entire 226 bp sequence. Each new SOX3_hs8A enhancer-test construct containing one of the 14 small deletions was injected into zebrafish embryos as done in previous experiments. The activities of these mutant versions were analyzed by monitoring GFP expression in both F0 and F1 generation. The correct expression pattern was affected in nearly all tested deletion cases, but the main domains were often kept at least in a weaker form. Three deletions, however, abolished the 2 dpf pattern completely and another three led to misregulation into other structures. Conserved TFBSs predictions that overlapped with the activating motifs were sites of zinc-finger protein, Pou3F2 and SRY- protein binding. This element suggests a modular type of functional organization containing several TFBSs with both strong and weak influences on the final expression pattern.

4. Discussion

4.1 The Essence of conserved synteny

Studies of chromosome evolution were for a long time dominated by the random breakage theory, which implies that there are no rearrangement hot spots in the human genome [193]. Based on the notion that any transformation of mouse gene order into human gene order would require a large number of breakpoint reuses Pevzner and Tesler rejected the random breakage model and proposed an alternative "fragile breakage" model of chromosome evolution [64, 194]. Instead of a biophysical constraint, a selection constraint causing non-random chromosome breakage was proposed. The link between rearrangements and regulatory regions of specific genes, which could not have been determined by computational means alone, became obvious to several independent research [25, 67, 68]. While these accounts appeared anecdotal, they were helped by additional evidence on a whole-genome scale. This was attained through our enhancer detection screen in zebrafish, revealing functional regulatory domains to span multiple unrelated genes corresponding to conserved syntenic regions. The evolutionarily distant human and zebrafish genomes share the shortest syntenic blocks, which can be thought to correspond to ancient essential cis-regulatory units. The ancestor of teleost fish experienced a whole genome duplication event and zebrafish has retained many duplicates. We understand evolutionary changes in HCNE and gene composition in these duplicated loci to be further proof of GRB arrangement in vertebrate genomes. The first step in the bioinformatic analysis of the cis-regulation is phylogenetic footprinting. Applying this method to human/zebrafish synteny blocks confirmed that there is high density of evolutionarily conserved elements around target genes and that the ancient chromosomal breakpoints do not occur in these HCNE dense regions keeping essential cis-regulatory information unchanged. The limitation of human/zebrafish comparisons is the small number of conserved sequences between these species that preclude the widespread use of this type of sequence comparison in many genomic regions.

As mentioned in section 1.3.3, co-regulation mechanism can underlie preservation of blocks of conserved synteny encompassing multiple target genes – these arised by tandem gene duplication form clusters of genes of which typical example are Hox genes. The conserved synteny in the *SOX11* region among all vertebrate genomes extends to *ID2*. *Sox11* and *Id2* have notably similar expression patterns indicating possible functional relation. Indeed, *Id2*, which functions as a passive repressor of proneural protein activity, represses the level of *Sox11* gene expression in chicken [195]. Therefore one cannot exclude the possibility that these two genes, although phylogenetically unrelated, are co-regulated by elements within their shared block of conserved synteny.

4.2 Transgenic zebrafish – a tool for testing of cis-regulatory sequences

In the mouse, putative regulatory elements are typically tested in a transient transgenic assay, using constructs with a minimal promoter and a lacZ reporter that becomes specifically activated by the element [70, 196]. Zebrafish GFP reporter assay has been used previously for this purpose also often in a transient expression assay. This assay is fast, but results in a high level of mosaicism and ectopic expression and limits the conclusions that can be drawn [30, 197]. In this thesis, I generated multiple transgenic lines for each element, precluding ambiguous results caused by position effect due to random integrations to the genome. This approach thus provides higher specificity and can reveal subtle details of expression patterns. In a set of tested elements from three loci (*sox3*, *PAX6* and *sox11*), we can conclude that the patterns are consistent with regulatory function for the assumed target gene in 60% of the elements. This number is higher than that resulting from the mouse assay [198] partly because of the larger window of developmental stages that can be analyzed (for mouse it is single embryonic time point) and because some enhancers show temporal specificity. This may underlie the inconclusive results for many of the remaining 40% of tested HCNEs that may act as regulators of very early or adult gene expression. Further negative results could be due to the requirement of some enhancers to act in combinations [197]. This could be tested in our assay only by coupling the elements together next to each other as the Tol2 system has certain sequence-length limitation and the natural distances are too long in both zebrafish and human genomes. Lastly, inconclusive results could be due to steric hindrance of the TFBSs due positioning relative to the promoter. Therefore, negative results reported in our experimental dataset (as well as those reported using the mouse) do not necessarily imply that this conserved element is not a transcriptional enhancer.

We assessed convenience of zebrafish as a model for testing human enhancers. We compared patterns of lacZ reporter expression in mouse driven by given human element to patterns obtained in the zebrafish GFP reporter assay using the same sequence. In all 5 cases the element activated reporter expression in comparable anatomical structures in mouse and zebrafish confriming the assumption, that sets of transcription factors and their binding properties remain conserved. Other proposition is that the evolution of developmental networks has its main source in the TFBS module mutation and not on the protein level.

Another idea that was to be tested was that regulatory elements define a GRB, and that appeared to hold true: we never observed reproducible expression patterns that were inconsistent with the expression of the target gene the genomic region. However, there is a number of complications that prevents smooth annotation of regulatory elements, namely position effects due to the random integration of the transgene. Interestingly, there were large variations in the susceptibility to position effects between specific tested sequences. One might hypothesize that this may be due to TFBSs affinity levels, the number of TFBs present in an element, or natural interactivity of the element. How such elements interact is however at present not known.

Multiple enhancers regulating fully or partially overlapping expression patterns have been identified for a variety of developmental control genes e.g. [70, 170]. All three sets of enhancers (from *sox3*, *sox11* and *PAX6* GRBs) discovered in our screen exhibited this feature. This suggests that maintenance of enhancers with overlapping functions regulating developmental control genes is a general feature. It would allow changes in expression pattern to arise from mutations that alter regulatory activity while preserving the required gene function [199]. The simultaneous activity of many enhancers in the same domain may be important as a backup mechanism that guarantees continuous expression during development and continued expression even upon inactivation or deletion of a particular regulatory region. We did not, however perform detailed sequence analyses to reveal redundancy on the TFBSs level, as reported in the SOX10 case [70].

4.3 Functional nature of HCNEs

As it has been established by previous research e.g.: [30, 69, 74, 164] high levels of non-protein coding evolutionary conservation coincide with important regulatory functions of these elements. Generally, one could imply that these elements function as TFBS modules. This should however not be over-generalized. As mentioned previously, there is a significant proportion of HCNEs that could not be assigned this function using the experimental approaches by the ENCODE consortium [18]. Further, a number of elements did not result in conclusive patterns in reporter assays including some in this thesis. A number of the HCNEs could encode as yet undetected non-coding RNA structures, cryptic promoters or un-annotated exons. HCNEs were previously demonstrated to encompass elements that regulate constitutive and alternative pre-mRNA splicing [200], and matrix attachment regions [201], promoter targeting sequences [142] or insulators [123, 202].

The mechanism(s) of conservation remains unknown. It has been speculated that HCNEs are mutational cold spots or regions where every site is under weak but still detectable negative selection. Others propose that the function in gene regulation is only one aspect of an HCNE; an additional role could lie in homologous chromosome pairing or other unknown processes adding constraint on the sequence conservation.

4.4 Modular nature of enhancers

Enhancers are highly structured with precisely arranged transcription factor binding sites that assemble so-called enhanceosomes of cooperating proteins [203]. On average, such modules contain 6–15 binding sites that bind four to eight different transcription factors [40]. To test how many, and which, functional TFBS units combine to form a single module, I designed an enhancer-deletion experiment on HCNE SOX3_hs8A. A series of TFBS-prediction-targeted deletions were introduced into the 226 bp sequence that was previously tested as a *SOX3*-specific enhancer. Changes in the original expression pattern were frequent – nearly all versions influenced either the expression level, or specificity. This suggested possible activating interactions of DNA binding by zinc-finger proteins, SRY-proteins, and Pou3F2 factor in two of the deleted sites. Pou3F2 is known to act cooperatively with the SOXB1/C factors and *SOX2* was shown to be activated by members of the POU family [204, 205]. As expression patterns of two predicted factors (heart-specific Nkx2-5 or somite-specific Myf) do not overlap with the *SOX3* pattern, we can likely exclude them from further speculations.

As we observed differences between the human and zebrafish expression pattern driven by the element #8 , using human-mouse alignment to predict mammal-specific factors could bring more relevant predictions. However, even for such a short sequence there are too many predictions precluding this attempt (52 predictions of conserved TFBSs by Transfac v10.0 for the 226 bp long alignment).

A detrimental effect on the module function could also have occurred by changes in the spacing of TFBSs or displacement of a TFBS from a favorable position (e.g. minor groove) on the DNA by deletion.

4.5 Enhancer-promoter specificity

The large size of the regulatory domain of one gene and interdigitation of its enhancers with coding regions of other,- apparently unaffected-, genes raises the question, why these genes do not respond to a proximal signal. One of the hypotheses (introduced in the section 1.4) is that functional motifs in the core promoter sequence underlie the target selection of a specific enhancer. Studies in Drosophila revealed transcriptional enhancers that are specific for promoters that contain either DPE or TATA box elements [206]. The polymerase stalling state was also proposed to be associated with certain promoter motifs [85].

By our results in the zebrafish transgenic GFP assay, the enhancer SOX3_hs8A specifically activated all promoters tested. Selected human and zebrafish promoter sequences were both from single tissue specific and developmental genes, of various categories by the criteria of Carnici and colleagues [87] and also by the target/bystander gene definition. As positive and negative controls, we used the endogenous *SOX3* human promoter to confirm the pattern by natural combination and SOX3_hs8A by itself to show that this promoter has no specific properties. Pattern differences observed were only in the level of expression and minor pattern variations perhaps caused by proximal elements included into the selected minimal promoter sequence.

To assess the presence and activity of putative insulators between the *SOX3* and *ATP11C* gene, we searched the ChIP-chip data by Kim et al. [95]. The only CTCF binding signal between the genes was found 445 kb from *SOX3* and 225 kb from *ATP11C* gene and 30 kb from the SOX3_hs8 enhancer. This sequence, however, did not impede promoter-enhancer interactions in the transgenic assay when inserted between enhancer and promoter.

CTCF is a protein highly conserved even between human and zebrafish, (up to 98% identity in the zinc finger region amino acid sequence) [207], but for zebrafish no functional data are available. CTCF binding sites occur frequently in genomes of all eutherian mammals, as well as opossum, chicken, and the pufferfish Tetraodon.

The motif shows a similar total number of instances across all vertebrate species despite a 5-fold variation in genome size. This is consistent with the motif being related to gene number (which is fairly constant across these species) rather than genome size [123]. Still, we cannot exclude that the insulator function known in the human genome is not conserved, but much more plausible explanation of the result of our insulator-test assay is in indications that CTCF has a more general role in genomic organization depending on the cofactors recruited to its binding sites [208].

Both results from promoter-interaction and insulator-test experiments support the idea that promoter-enhancer specificity is mediated by promoter-tethering elements [209], promoter mutual competition, insulators, or by cryptic promoters [210]. Epigenetic changes of promoter availability are another possible mechanism.

4.6 Subfunctionalization apparent on *sox11* regulatory elements

Following the whole genome duplication, the expression profiles of retained gene duplicates diverge. To examine divergence process at the regulation level of these genes, we examined expression pattern of dissected cis-regulatory information of zebrafish *sox11a/b*. *Sox11* belongs to the genes predisposed to the subfunctionalization process as it has high level 'and wide breadth of expression [211]. We can also explain this predisposition by the presence of multiple cis-regulatory modules offering space for neutral evolution through both degenerative and compensatory mutations without fatal consequences for the gene function.

The results in this thesis, together with other recent reports [212], provide experimental support that differential divergence of HCNEs of paralogs may be a general phenomenon in vertebrates according to the subfunctionalization (or duplication-degeneration-complementation) model. The coexistence of subfunctionalization and neofunctionalization together are another discussed alternative. In our study we lack data to conclude for the neofunctionalization process, as we could not compare numbers of verified functional motifs common vs. different between the two paralogs. Moreover, there is no ancestral expression data that would allow us to conclude for new functions within the elements. Previous analyses on yeast [213, 214] reveal that with increasing time, subfunctionalization decreases in importance and its role seems to be to preserve duplicate copies for eventual neofunctionalization, a role as a transition state. However, the relative roles of subfunctionalization and neofunctionalization in the retention of duplicate genes remain to be clarified, especially for higher eukaryotes.

4.7 New candidate genes for genome wide association studies

Genome-wide association studies typically aim to find protein-coding alleles that explain a given trait or disease. Type 2 diabetes is one of the complex human diseases, whereas for quantitative traits, the gene expression level alteration should be considered as a predisposing factor, which is often caused by regulatory variations [215].

To provide functional evidence for our hypothesis that many of the T2D-associated SNPs are in genomic regions devoted to long-range regulation of a developmental gene (the GRB target), we examined the general landscape around reported SNPs and confirmed, using a GFP-reporter assay in zebrafish, HCNEs that overlap or neighbor a subset of them to drive expression consistent with that of the GRB target gene. The expression patterns driven by human HCNEs near the risk SNPs support our hypothesis that the elements in question do not necessarily regulate the closest gene, but rather the GRB target.

The challenge in localizing regulatory polymorphisms is isolating the causal variants that are in linkage disequilibrium with many other variants. The particular SNPs chosen for each case in our study do not necessarily represent a regulatory mutation, but instead one might need to examine more putative regulatory activities encoded within the haplotype region. It would be relatively simple to carry out a systematic analysis to search for enhancer functions in the region which best fit the disease phenotype and the causative SNPs within them. However, interpreting a specific consequence of a mutation in such enhancer could be difficult by our reporter assay in cases where the change in the pattern is quantitative and subtle. The combination of target gene-microarray techniques for phenotyping and linkage analyses combined with SNP association would be needed to narrow down the candidate regulatory determinants that contribute to variation in target gene expression. Possible qualitative changes by the regulatory mutation as those observed previously [167] would provide strong support for regulatory mutations involved in pathogenic processes.

It is expected that with growing numbers of identified human enhancers it will become possible to target systematic screens for regulatory mutations in the distant-acting class of gene regulatory elements and with the knowledge of genome architecture also to prevent incorrect gene assignment.

5. List of abbreviations

BAC - bacterial artificial chromosome

bp –base pair

ChIP – chromatin immunoprecipitation

DPE – downstream promoter element

DNA – deoxyribonucleic acid

ENCODE – encyclopaedia of DNA elements

EST – expressed sequence tag

GRB – genomic regulatory block

HCNE – highly conserved non-coding element

RNA – ribonucleic acid

SNP – single nucleotide polymorphism

T2D – Type-two diabetes

TFBS – transcription factor binding site

TF – transcription factor

TSS – transcription start site

UCR – ultraconserved region

UCSC – University of California Santa Cruz

6. References

1. Cretekos, C.J., et al., *Regulatory divergence modifies limb length between mammals.* Genes Dev, 2008. **22**(2): p. 141-51.
2. Levine, M. and R. Tjian, *Transcription regulation and animal diversity.* Nature, 2003. **424**(6945): p. 147-51.
3. Lindblad-Toh, K., et al., *Genome sequence, comparative analysis and haplotype structure of the domestic dog.* Nature, 2005. **438**(7069): p. 803-19.
4. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
5. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
6. Jekosch, K., *The zebrafish genome project: sequence analysis and annotation.* Methods Cell Biol, 2004. **77**: p. 225-39.
7. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.
8. Margulies, E.H. and E. Birney, *Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes.* Nat Rev Genet, 2008. **9**(4): p. 303-13.
9. Adams, M.D., et al., *The genome sequence of Drosophila melanogaster.* Science, 2000. **287**(5461): p. 2185-95.
10. Wood, V., et al., *The genome sequence of Schizosaccharomyces pombe.* Nature, 2002. **415**(6874): p. 871-80.
11. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D773-9.
12. Flicek, P., et al., *Ensembl 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D707-14.
13. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2008. **36**(Database issue): p. D13-21.
14. Clamp, M., et al., *Distinguishing protein-coding and noncoding genes in the human genome.* Proc Natl Acad Sci U S A, 2007. **104**(49): p. 19428-33.
15. Harrow, J., et al., *GENCODE: producing a reference annotation for ENCODE.* Genome Biol, 2006. **7 Suppl 1**: p. S4 1-9.
16. Denoeud, F., et al., *Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions.* Genome Res, 2007. **17**(6): p. 746-59.
17. Pedersen, J.S., et al., *Identification and classification of conserved RNA secondary structures in the human genome.* PLoS Comput Biol, 2006. **2**(4): p. e33.
18. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
19. Brown, C.J., et al., *A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome.* Nature, 1991. **349**(6304): p. 38-44.
20. Duret, L., et al., *The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene.* Science, 2006. **312**(5780): p. 1653-5.
21. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs.* Genome Res, 2007. **17**(5): p. 556-65.

22. Gerstein, M.B., et al., *What is a gene, post-ENCODE? History and updated definition.* Genome Res, 2007. **17**(6): p. 669-81.
23. Kimura, M., *The Neutral Theory of Molecular Evolution*. 1983, Cambridge: Cambridge University Press.
24. Elgar, G., et al., *Small is beautiful: comparative genomics with the pufferfish (Fugu rubripes).* Trends Genet, 1996. **12**(4): p. 145-50.
25. Gomez-Skarmeta, J.L., B. Lenhard, and T.S. Becker, *New technologies, new findings, and new concepts in the study of vertebrate cis-regulatory sequences.* Dev Dyn, 2006. **235**(4): p. 870-85.
26. Wasserman, W.W., et al., *Human-mouse genome comparisons to locate regulatory sites.* Nat Genet, 2000. **26**(2): p. 225-8.
27. Kumar, S. and S.B. Hedges, *A molecular timescale for vertebrate evolution.* Nature, 1998. **392**(6679): p. 917-20.
28. Engstrom, P.G., D. Fredman, and B. Lenhard, *Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes.* Genome Biol, 2008. **9**(2): p. R34.
29. Aparicio, S., et al., *Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes.* Science, 2002. **297**(5585): p. 1301-10.
30. Woolfe, A., et al., *Highly conserved non-coding sequences are associated with vertebrate development.* PLoS Biol, 2005. **3**(1): p. e7.
31. Kammandel, B., et al., *Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity.* Dev Biol, 1999. **205**(1): p. 79-97.
32. Sandelin, A., et al., *Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes.* BMC Genomics, 2004. **5**(1): p. 99.
33. Dermitzakis, E.T. and A.G. Clark, *Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover.* Mol Biol Evol, 2002. **19**(7): p. 1114-21.
34. Sironi, M., et al., *Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences.* Hum Mol Genet, 2005. **14**(17): p. 2533-46.
35. Sanges, R., et al., *Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage.* Genome Biol, 2006. **7**(7): p. R56.
36. Lettice, L.A., et al., *Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly.* Proc Natl Acad Sci U S A, 2002. **99**(11): p. 7548-53.
37. Simons, C., et al., *Transposon-free regions in mammalian genomes.* Genome Res, 2006. **16**(2): p. 164-72.
38. Simons, C., et al., *Maintenance of transposon-free regions throughout vertebrate evolution.* BMC Genomics, 2007. **8**: p. 470.
39. Margulies, E.H., et al., *Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome.* Genome Res, 2007. **17**(6): p. 760-74.
40. Wray, G.A., et al., *The evolution of transcriptional regulation in eukaryotes.* Mol Biol Evol, 2003. **20**(9): p. 1377-419.
41. Hoegg, S., et al., *Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish.* J Mol Evol, 2004. **59**(2): p. 190-203.
42. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.
43. Kleinjan, D.A., et al., *Subfunctionalization of Duplicated Zebrafish pax6 Genes by cis-Regulatory Divergence.* PLoS Genet, 2008. **4**(2): p. e29.

44. Vavouri, T., et al., *Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans.* Genome Biol, 2007. **8**(2): p. R15.

45. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.

46. McEwen, G.K., et al., *Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis.* Genome Res, 2006. **16**(4): p. 451-65.

47. Bejerano, G., et al., *A distal enhancer and an ultraconserved exon are derived from a novel retroposon.* Nature, 2006. **441**(7089): p. 87-90.

48. Lowe, C.B., G. Bejerano, and D. Haussler, *Thousands of human mobile element fragments undergo strong purifying selection near developmental genes.* Proc Natl Acad Sci U S A, 2007. **104**(19): p. 8005-10.

49. Xie, X., M. Kamal, and E.S. Lander, *A family of conserved noncoding elements derived from an ancient transposable element.* Proc Natl Acad Sci U S A, 2006. **103**(31): p. 11659-64.

50. Kapitonov, V.V. and J. Jurka, *A novel class of SINE elements derived from 5S rRNA.* Mol Biol Evol, 2003. **20**(5): p. 694-702.

51. Kamal, M., X. Xie, and E.S. Lander, *A large family of ancient repeat elements in the human genome is under strong selection.* Proc Natl Acad Sci U S A, 2006. **103**(8): p. 2740-5.

52. Sasaki, T., et al., *Possible involvement of SINEs in mammalian-specific brain formation.* Proc Natl Acad Sci U S A, 2008. **105**(11): p. 4220-5.

53. Gould, S.J. and E.S. Vrba, *Exaptation: A missing term in the science of form.* Paleobiology, 1982. **8**(1): p. 4-15.

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

55. Tuch, B.B., H. Li, and A.D. Johnson, *Evolution of eukaryotic transcription circuits.* Science, 2008. **319**(5871): p. 1797-9.

56. Bejerano, G., et al., *Ultraconserved elements in the human genome.* Science, 2004. **304**(5675): p. 1321-5.

57. Kryukov, G.V., S. Schmidt, and S. Sunyaev, *Small fitness effect of mutations in highly conserved non-coding regions.* Hum Mol Genet, 2005. **14**(15): p. 2221-9.

58. Chen, C.T., J.C. Wang, and B.A. Cohen, *The strength of selection on ultraconserved elements in the human genome.* Am J Hum Genet, 2007. **80**(4): p. 692-704.

59. Katzman, S., et al., *Human genome ultraconserved elements are ultraselected.* Science, 2007. **317**(5840): p. 915.

60. Ahituv, N., et al., *Deletion of ultraconserved elements yields viable mice.* PLoS Biol, 2007. **5**(9): p. e234.

61. Visel, A., et al., *Ultraconservation identifies a small subset of extremely constrained developmental enhancers.* Nat Genet, 2008. **40**(2): p. 158-60.

62. Ono, S., *Ancient linkage groups and frozen accidents.* Nature, 1973. **244**(5414): p. 259-62.

63. Bourque, G., P.A. Pevzner, and G. Tesler, *Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes.* Genome Res, 2004. **14**(4): p. 507-16.

64. Pevzner, P. and G. Tesler, *Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7672-7.

65. Becker, T.S. and B. Lenhard, *The random versus fragile breakage models of chromosome evolution: a matter of resolution.* Mol Genet Genomics, 2007. **278**(5): p. 487-91.
66. Engstrom, P.G., et al., *Genomic regulatory blocks underlie extensive microsynteny conservation in insects.* Genome Res, 2007. **17**(12): p. 1898-908.
67. Ahituv, N., et al., *Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny.* Hum Mol Genet, 2005. **14**(20): p. 3057-63.
68. Mackenzie, A., K.A. Miller, and J.M. Collinson, *Is there a functional link between gene interdigitation and multi-species conservation of synteny blocks?* Bioessays, 2004. **26**(11): p. 1217-24.
69. Nobrega, M.A., et al., *Scanning human gene deserts for long-range enhancers.* Science, 2003. **302**(5644): p. 413.
70. Werner, T., et al., *Multiple conserved regulatory elements with overlapping functions determine Sox10 expression in mouse embryogenesis.* Nucleic Acids Res, 2007. **35**(19): p. 6526-38.
71. Nobrega, M.A., et al., *Megabase deletions of gene deserts result in viable mice.* Nature, 2004. **431**(7011): p. 988-93.
72. Kikuta, H., et al., *Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates.* Genome Res, 2007. **17**(5): p. 545-55.
73. Kikuta, H., et al., *Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes.* Genome Biol, 2007. **8 Suppl 1**: p. S4.
74. de la Calle-Mustienes, E., et al., *A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts.* Genome Res, 2005. **15**(8): p. 1061-72.
75. Spitz, F., F. Gonzalez, and D. Duboule, *A global control region defines a chromosomal regulatory landscape containing the HoxD cluster.* Cell, 2003. **113**(3): p. 405-17.
76. Caron, H., et al., *The human transcriptome map: clustering of highly expressed genes in chromosomal domains.* Science, 2001. **291**(5507): p. 1289-92.
77. Gierman, H.J., et al., *Domain-wide regulation of gene expression in the human genome.* Genome Res, 2007. **17**(9): p. 1286-95.
78. Lercher, M.J., A.O. Urrutia, and L.D. Hurst, *Clustering of housekeeping genes provides a unified model of gene order in the human genome.* Nat Genet, 2002. **31**(2): p. 180-3.
79. Singer, G.A., et al., *Clusters of co-expressed genes in mammalian genomes are conserved by natural selection.* Mol Biol Evol, 2005. **22**(3): p. 767-75.
80. Spitz, F. and D. Duboule, *Chapter 6 global control regions and regulatory landscapes in vertebrate development and evolution.* Adv Genet, 2008. **61**: p. 175-205.
81. Core, L.J. and J.T. Lis, *Transcription regulation through promoter-proximal pausing of RNA polymerase II.* Science, 2008. **319**(5871): p. 1791-2.
82. Kong, S., et al., *Transcription of the HS2 enhancer toward a cis-linked gene is independent of the orientation, position, and distance of the enhancer relative to the gene.* Mol Cell Biol, 1997. **17**(7): p. 3955-65.
83. Gaszner, M. and G. Felsenfeld, *Insulators: exploiting transcriptional and epigenetic mechanisms.* Nat Rev Genet, 2006. **7**(9): p. 703-13.
84. Fuss, S.H., M. Omura, and P. Mombaerts, *Local and cis effects of the H element on expression of odorant receptor genes in mouse.* Cell, 2007. **130**(2): p. 373-84.

85. Zeitlinger, J., et al., *RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo.* Nat Genet, 2007. **39**(12): p. 1512-6.

86. Sandelin, A., et al., *Mammalian RNA polymerase II core promoters: insights from genome-wide studies.* Nat Rev Genet, 2007. **8**(6): p. 424-36.

87. Carninci, P., et al., *Genome-wide analysis of mammalian promoter architecture and evolution.* Nat Genet, 2006. **38**(6): p. 626-35.

88. Dekker, J., *Gene regulation in the third dimension.* Science, 2008. **319**(5871): p. 1793-4.

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

90. Picker, A., et al., *A novel positive transcriptional feedback loop in midbrain-hindbrain boundary development is revealed through analysis of the zebrafish pax2.1 promoter in transgenic lines.* Development, 2002. **129**(13): p. 3227-39.

91. Du, S.J. and M. Dienhart, *Zebrafish tiggy-winkle hedgehog promoter directs notochord and floor plate green fluorescence protein expression in transgenic zebrafish embryos.* Dev Dyn, 2001. **222**(4): p. 655-66.

92. Chen, Y.H., et al., *Multiple upstream modules regulate zebrafish myf5 expression.* BMC Dev Biol, 2007. **7**: p. 1.

93. Ellingsen, S., et al., *Large-scale enhancer detection in the zebrafish genome.* Development, 2005. **132**(17): p. 3799-811.

94. Cai, H. and M. Levine, *Modulation of enhancer-promoter interactions by insulators in the Drosophila embryo.* Nature, 1995. **376**(6540): p. 533-6.

95. Kim, T.H., et al., *Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome.* Cell, 2007. **128**(6): p. 1231-45.

96. Wendt, K.S., et al., *Cohesin mediates transcriptional insulation by CCCTC-binding factor.* Nature, 2008. **451**(7180): p. 796-801.

97. Parelho, V., et al., *Cohesins functionally associate with CTCF on mammalian chromosome arms.* Cell, 2008. **132**(3): p. 422-33.

98. Rubio, E.D., et al., *CTCF physically links cohesin to chromatin.* Proc Natl Acad Sci U S A, 2008.

99. Li, Q., G. Barkess, and H. Qian, *Chromatin looping and the probability of transcription.* Trends Genet, 2006. **22**(4): p. 197-202.

100. Cai, S., C.C. Lee, and T. Kohwi-Shigematsu, *SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes.* Nat Genet, 2006. **38**(11): p. 1278-88.

101. Wang, Q., J.S. Carroll, and M. Brown, *Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking.* Mol Cell, 2005. **19**(5): p. 631-42.

102. Ling, J., et al., *The HS2 enhancer of the beta-globin locus control region initiates synthesis of non-coding, polyadenylated RNAs independent of a cis-linked globin promoter.* J Mol Biol, 2005. **350**(5): p. 883-96.

103. Zhu, X., et al., *A facilitated tracking and transcription mechanism of long-range enhancer function.* Nucleic Acids Res, 2007. **35**(16): p. 5532-44.

104. Chuang, C.H., et al., *Long-range directional movement of an interphase chromosome site.* Curr Biol, 2006. **16**(8): p. 825-31.

105. Dundr, M., et al., *Actin-dependent intranuclear repositioning of an active gene locus in vivo.* J Cell Biol, 2007. **179**(6): p. 1095-103.

106. Spilianakis, C.G., et al., *Interchromosomal associations between alternatively expressed loci.* Nature, 2005. **435**(7042): p. 637-45.

107. Gohl, D., et al., *Enhancer blocking and transvection at the Drosophila apterous locus.* Genetics, 2008. **178**(1): p. 127-43.
108. Kumaran, R.I., R. Thakar, and D.L. Spector, *Chromatin dynamics and gene positioning.* Cell, 2008. **132**(6): p. 929-34.
109. Small, K.S., et al., *A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome.* Genome Biol, 2007. **8**(3): p. R41.
110. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.
111. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes Dev, 2002. **16**(1): p. 6-21.
112. Bernstein, B.E., A. Meissner, and E.S. Lander, *The mammalian epigenome.* Cell, 2007. **128**(4): p. 669-81.
113. Bock, C., et al., *CpG island mapping by epigenome prediction.* PLoS Comput Biol, 2007. **3**(6): p. e110.
114. Tanay, A., et al., *Hyperconserved CpG domains underlie Polycomb-binding sites.* Proc Natl Acad Sci U S A, 2007. **104**(13): p. 5521-6.
115. Lee, T.I., et al., *Control of developmental regulators by Polycomb in human embryonic stem cells.* Cell, 2006. **125**(2): p. 301-13.
116. Ringrose, L., *Polycomb comes of age: genome-wide profiling of target sites.* Curr Opin Cell Biol, 2007. **19**(3): p. 290-7.
117. Schones, D.E. and K. Zhao, *Genome-wide approaches to studying chromatin modifications.* Nat Rev Genet, 2008. **9**(3): p. 179-91.
118. Lande-Diner, L., et al., *Role of DNA methylation in stable gene repression.* J Biol Chem, 2007. **282**(16): p. 12194-200.
119. McGaughey, D.M., et al., *Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b.* Genome Res, 2008. **18**(2): p. 252-60.
120. Uchikawa, M., et al., *Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals.* Dev Cell, 2003. **4**(4): p. 509-19.
121. Fisher, S., et al., *Conservation of RET regulatory function from human to zebrafish without sequence similarity.* Science, 2006. **312**(5771): p. 276-9.
122. Boffelli, D., et al., *Phylogenetic shadowing of primate sequences to find functional regions of the human genome.* Science, 2003. **299**(5611): p. 1391-4.
123. Xie, X., et al., *Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.* Proc Natl Acad Sci U S A, 2007. **104**(17): p. 7145-50.
124. Segal, E., et al., *Predicting expression patterns from regulatory sequence in Drosophila segmentation.* Nature, 2008. **451**(7178): p. 535-40.
125. Loots, G.G., *Chapter 10 genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis.* Adv Genet, 2008. **61**: p. 269-93.
126. Miller, W., et al., *28-way vertebrate alignment and conservation track in the UCSC Genome Browser.* Genome Res, 2007. **17**(12): p. 1797-808.
127. Bray, N. and L. Pachter, *MAVID multiple alignment server.* Nucleic Acids Res, 2003. **31**(13): p. 3525-6.
128. Brudno, M., et al., *LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.* Genome Res, 2003. **13**(4): p. 721-31.
129. Dubchak, I., *Comparative analysis and visualization of genomic sequences using VISTA browser and associated computational tools.* Methods Mol Biol, 2007. **395**: p. 3-16.

130. Ovcharenko, I., et al., *ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W280-6.

131. Sandelin, A., W.W. Wasserman, and B. Lenhard, *ConSite: web-based prediction of regulatory elements using cross-species comparison.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W249-52.

132. Garner, M.M. and A. Revzin, *A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system.* Nucleic Acids Res, 1981. **9**(13): p. 3047-60.

133. Ellington, A.D. and J.W. Szostak, *In vitro selection of RNA molecules that bind specific ligands.* Nature, 1990. **346**(6287): p. 818-22.

134. Wu, C., *The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I.* Nature, 1980. **286**(5776): p. 854-60.

135. Crawford, G.E., et al., *Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites.* Proc Natl Acad Sci U S A, 2004. **101**(4): p. 992-7.

136. Crawford, G.E., et al., *DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays.* Nat Methods, 2006. **3**(7): p. 503-9.

137. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.

138. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.

139. van Steensel, B. and S. Henikoff, *Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase.* Nat Biotechnol, 2000. **18**(4): p. 424-8.

140. Simonis, M., J. Kooren, and W. de Laat, *An evaluation of 3C-based methods to capture DNA interactions.* Nat Methods, 2007. **4**(11): p. 895-901.

141. Carter, D., et al., *Long-range chromatin regulatory interactions in vivo.* Nat Genet, 2002. **32**(4): p. 623-6.

142. Zhou, J. and M. Levine, *A novel cis-regulatory element, the PTS, mediates an anti-insulator activity in the Drosophila embryo.* Cell, 1999. **99**(6): p. 567-75.

143. Pirrotta, V., H. Steller, and M.P. Bozzetti, *Multiple upstream regulatory elements control the expression of the Drosophila white gene.* Embo J, 1985. **4**(13A): p. 3501-8.

144. O'Kane, C.J. and W.J. Gehring, *Detection in situ of genomic regulatory elements in Drosophila.* Proc Natl Acad Sci U S A, 1987. **84**(24): p. 9123-7.

145. Halfon, M.S., S.M. Gallo, and C.M. Bergman, *REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila.* Nucleic Acids Res, 2008. **36**(Database issue): p. D594-8.

146. Dehal, P., et al., *The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins.* Science, 2002. **298**(5601): p. 2157-67.

147. Imai, K.S., et al., *Regulatory blueprint for a chordate embryo.* Science, 2006. **312**(5777): p. 1183-7.

148. Brown, C.D., D.S. Johnson, and A. Sidow, *Functional architecture and evolution of transcriptional elements that drive gene coexpression.* Science, 2007. **317**(5844): p. 1557-60.

149. Sasakura, Y., et al., *Enhancer detection in the ascidian Ciona intestinalis with transposase-expressing lines of Minos.* Dev Dyn, 2008. **237**(1): p. 39-50.

150. *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.* Nature, 2004. **432**(7018): p. 695-716.

151. Sato, Y., et al., *Stable integration and conditional expression of electroporated transgenes in chicken embryos.* Dev Biol, 2007. **305**(2): p. 616-24.
152. Allen, N.D., et al., *Transgenes as probes for active chromosomal domains in mouse development.* Nature, 1988. **333**(6176): p. 852-5.
153. Korn, R., et al., *Enhancer trap integrations in mouse embryonic stem cells give rise to staining patterns in chimaeric embryos with a high frequency and detect endogenous genes.* Mech Dev, 1992. **39**(1-2): p. 95-109.
154. Kothary, R., et al., *A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube.* Nature, 1988. **335**(6189): p. 435-7.
155. Pennacchio, L.A., et al., *In vivo enhancer analysis of human conserved non-coding sequences.* Nature, 2006. **444**(7118): p. 499-502.
156. Kroll, K.L. and E. Amaya, *Transgenic Xenopus embryos from sperm nuclear transplantations reveal FGF signaling requirements during gastrulation.* Development, 1996. **122**(10): p. 3173-83.
157. Ogino, H., W.B. McConnell, and R.M. Grainger, *Highly efficient transgenesis in Xenopus tropicalis using I-SceI meganuclease.* Mech Dev, 2006. **123**(2): p. 103-13.
158. Hamlet, M.R., et al., *Tol2 transposon-mediated transgenesis in Xenopus tropicalis.* Genesis, 2006. **44**(9): p. 438-45.
159. Allen, B.G. and D.L. Weeks, *Using phiC31 integrase to make transgenic Xenopus laevis embryos.* Nat Protoc, 2006. **1**(3): p. 1248-57.
160. Laplante, M., et al., *Enhancer detection in the zebrafish using pseudotyped murine retroviruses.* Methods, 2006. **39**(3): p. 189-98.
161. Koga, A., et al., *Transposable element in fish.* Nature, 1996. **383**(6595): p. 30.
162. Kawakami, K., *Tol2: a versatile gene transfer vector in vertebrates.* Genome Biol, 2007. **8 Suppl 1**: p. S7.
163. Yang, Z., et al., *Modified bacterial artificial chromosomes for zebrafish transgenesis.* Methods, 2006. **39**(3): p. 183-8.
164. Conte, I. and P. Bovolenta, *Comprehensive characterization of the cis-regulatory code responsible for the spatio-temporal expression of olSix3.2 in the developing medaka forebrain.* Genome Biol, 2007. **8**(7): p. R137.
165. Clark, R.M., et al., *Reciprocal mouse and human limb phenotypes caused by gain- and loss-of-function mutations affecting Lmbr1.* Genetics, 2001. **159**(2): p. 715-26.
166. Lettice, L.A., et al., *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.* Hum Mol Genet, 2003. **12**(14): p. 1725-35.
167. Lettice, L.A., et al., *Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly.* Hum Mol Genet, 2008. **17**(7): p. 978-85.
168. Sagai, T., et al., *Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb.* Development, 2005. **132**(4): p. 797-803.
169. Yakushiji, N., et al., *Correlation between Shh expression and DNA methylation status of the limb-specific Shh enhancer region during limb regeneration in amphibians.* Dev Biol, 2007. **312**(1): p. 171-82.
170. Jeong, Y., et al., *A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers.* Development, 2006. **133**(4): p. 761-72.
171. Glaser, T., et al., *PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects.* Nat Genet, 1994. **7**(4): p. 463-71.

172. Schedl, A., et al., *Influence of PAX6 gene dosage on development: overexpression causes severe eye abnormalities.* Cell, 1996. **86**(1): p. 71-82.
173. Fantes, J., et al., *Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype.* Hum Mol Genet, 1995. **4**(3): p. 415-22.
174. Lauderdale, J.D., et al., *3' deletions cause aniridia by preventing PAX6 gene expression.* Proc Natl Acad Sci U S A, 2000. **97**(25): p. 13755-9.
175. Kleinjan, D.A., et al., *Long-range downstream enhancers are essential for Pax6 expression.* Dev Biol, 2006. **299**(2): p. 563-81.
176. Francke, U., et al., *Aniridia-Wilms' tumor association: evidence for specific deletion of 11p13.* Cytogenet Cell Genet, 1979. **24**(3): p. 185-92.
177. Kent, J., et al., *The reticulocalbin gene maps to the WAGR region in human and to the Small eye Harwell deletion in mouse.* Genomics, 1997. **42**(2): p. 260-7.
178. Bowl, M.R., et al., *An interstitial deletion-insertion involving chromosomes 2p25.3 and Xq27.1, near SOX3, causes X-linked recessive hypoparathyroidism.* J Clin Invest, 2005. **115**(10): p. 2822-31.
179. Kleinjan, D.A. and L.A. Lettice, *Chapter 13 long-range gene control and genetic disease.* Adv Genet, 2008. **61**: p. 339-88.
180. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
181. Scott, L.J., et al., *A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants.* Science, 2007.
182. Yeager, M., et al., *Genome-wide association study of prostate cancer identifies a second risk locus at 8q24.* Nat Genet, 2007. **39**(5): p. 645-649.
183. Shifman, S., et al., *Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women.* PLoS Genet, 2008. **4**(2): p. e28.
184. Zeggini, E., et al., *Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes.* Science, 2007. **316**(5829): p. 1336-41.
185. Saxena, R., et al., *Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.* Science, 2007. **316**(5829): p. 1331-6.
186. Steinthorsdottir, V., et al., *A variant in CDKAL1 influences insulin response and risk of type 2 diabetes.* Nat Genet, 2007.
187. Wei, F.Y., et al., *Cdk5-dependent regulation of glucose-stimulated insulin secretion.* Nat Med, 2005. **11**(10): p. 1104-8.
188. Wilson, M.E., et al., *The HMG box transcription factor Sox4 contributes to the development of the endocrine pancreas.* Diabetes, 2005. **54**(12): p. 3402-9.
189. Frayling, T.M., et al., *A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity.* Science, 2007.
190. Kobayashi, D., et al., *Early subdivisions in the neural plate define distinct competence for inductive signals.* Development, 2002. **129**(1): p. 83-93.
191. Karlsson, E.K., et al., *Efficient mapping of mendelian traits in dogs through genome-wide association.* Nat Genet, 2007. **39**(11): p. 1321-8.
192. Goldsworthy, M., et al., *The role of the transcription factor Sox4 in insulin secretion and impaired glucose tolerance.* Diabetes, 2008.
193. Nadeau, J.H. and B.A. Taylor, *Lengths of chromosomal segments conserved since divergence of man and mouse.* Proc Natl Acad Sci U S A, 1984. **81**(3): p. 814-8.
194. Peng, Q., P.A. Pevzner, and G. Tesler, *The fragile breakage versus random breakage models of chromosome evolution.* PLoS Comput Biol, 2006. **2**(2): p. e14.

195. Bergsland, M., et al., *The establishment of neuronal properties is controlled by Sox4 and Sox11.* Genes Dev, 2006. **20**(24): p. 3475-86.

196. Beermann, F., et al., *Identification of evolutionarily conserved regulatory elements in the mouse Fgf8 locus.* Genesis, 2006. **44**(1): p. 1-6.

197. Ertzer, R., et al., *Cooperation of sonic hedgehog enhancers in midline expression.* Dev Biol, 2007. **301**(2): p. 578-89.

198. Visel, A., et al., *VISTA Enhancer Browser--a database of tissue-specific human enhancers.* Nucleic Acids Res, 2007. **35**(Database issue): p. D88-92.

199. Carroll, S.B., *Endless forms: the evolution of gene regulation and morphological diversity.* Cell, 2000. **101**(6): p. 577-80.

200. Yeo, G.W., E.L. Nostrand, and T.Y. Liang, *Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements.* PLoS Genet, 2007. **3**(5): p. e85.

201. Glazko, G.V., et al., *A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions.* Trends Genet, 2003. **19**(3): p. 119-24.

202. Kmita, M., et al., *Evolutionary conserved sequences are required for the insulation of the vertebrate Hoxd complex in neural cells.* Development, 2002. **129**(23): p. 5521-8.

203. Panne, D., T. Maniatis, and S.C. Harrison, *An atomic model of the interferon-beta enhanceosome.* Cell, 2007. **129**(6): p. 1111-23.

204. Catena, R., et al., *Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells.* J Biol Chem, 2004. **279**(40): p. 41846-57.

205. Tanaka, S., et al., *Interplay of SOX and POU factors in regulation of the Nestin gene in neural primordial cells.* Mol Cell Biol, 2004. **24**(20): p. 8834-46.

206. Butler, J.E. and J.T. Kadonaga, *Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs.* Genes Dev, 2001. **15**(19): p. 2515-9.

207. Pugacheva, E.M., et al., *Cloning and characterization of zebrafish CTCF: Developmental expression patterns, regulation of the promoter region, and evolutionary aspects of gene organization.* Gene, 2006. **375**: p. 26-36.

208. Wallace, J.A. and G. Felsenfeld, *We gather together: insulators and genome organization.* Curr Opin Genet Dev, 2007. **17**(5): p. 400-7.

209. Calhoun, V.C., A. Stathopoulos, and M. Levine, *Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex.* Proc Natl Acad Sci U S A, 2002. **99**(14): p. 9243-7.

210. Carvajal, J.J., A. Keith, and P.W. Rigby, *Global transcriptional regulation of the locus encoding the skeletal muscle determination genes Mrf4 and Myf5.* Genes Dev, 2008. **22**(2): p. 265-76.

211. Semon, M. and K.H. Wolfe, *Preferential subfunctionalization of slow-evolving genes after allopolyploidization in Xenopus laevis.* Proc Natl Acad Sci U S A, 2008. **105**(24): p. 8333-8.

212. Hadzhiev, Y., et al., *Functional diversification of sonic hedgehog paralog enhancers identified by phylogenomic reconstruction.* Genome Biol, 2007. **8**(6): p. R106.

213. Papp, B., C. Pal, and L.D. Hurst, *Evolution of cis-regulatory elements in duplicated genes of yeast.* Trends Genet, 2003. **19**(8): p. 417-22.

214. Tirosh, I. and N. Barkai, *Comparative analysis indicates regulatory neofunctionalization of yeast duplicates.* Genome Biol, 2007. **8**(4): p. R50.

215. Morley, M., et al., *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. **430**(7001): p. 743-7.

216. Bryne, J.C., et al., *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D102-6.

217. Lenhard, B. and W.W. Wasserman, *TFBS: Computational framework for transcription factor binding site analysis.* Bioinformatics, 2002. **18**(8): p. 1135-6.