aus dem Institut für Biostatistik und Informatik in Medizin und Alternsforschung (IBIMA)
der Medizinischen Fakultät der Universität Rostock

Unraveling the Transcriptional *C*is-regulatory Code

kumulative
Habilitationsschrift
zur
Erlangung des akademischen Grades
doctor rerum naturalium habilitata (Dr. rer. nat. habil.)
der Universitätsmedizin Rostock

vorgelegt von

Leila Taher
aus Erlangen
geb. am 05. Mai 1978 in Santa Fe (Argentinien)

urn:nbn:de:gbv:28-diss2016-0113-4

Dekan Prof. Dr. med. Emil C. Reisinger
Universitätsmedizin Rostock

1. Gutachter Herr Prof. Dr. Wojciech Makalowski
Institut für Bioinformatik
Westfälische Wilhelms-Universität Münster

2. Gutachter Herr Prof. Dr. Burkhard Morgenstern
Institut für Mikrobiologie und Genetik
Georg-August-Universität Göttingen

3. Gutachter Herr Prof. Dr. Georg Füllen
Institut für Biostatistik und Informatik in Medizin und Alternsforschung
Universitätsmedizin Rostock

Datum der Einreichung: 30. April 2015
Datum der Verteidigung: 29. Februar 2016
Datum der Lehrprobe: 26. Mai 2016

Summary

Introduction

Starting with the Human Genome Project, genome research has had a fundamental impact on scientific progress. One of the surprises of the Human Genome Project was the relatively small number of protein-coding genes in the genome, estimated in roughly $23,000$. It is nowadays accepted that eukaryotic complexity is not dictated by the number of protein-coding genes of the genome, but rather achieved through the combinatorics of gene expression programs. Distinct aspects of the expression pattern of a gene are mediated by discrete regulatory sequences, known as *cis*-regulatory elements. *Cis*-regulatory elements are typically short and harbor binding sites for multiple transcription factors (TFs) in a particular arrangement, defining what we call *cis*-regulatory grammar. The advent of affordable high-throughput sequencing technologies has provided us with a plethora of genome-wide assays that have revolutionized our ability to interrogate the genome. Nevertheless, our understanding of gene regulation remains incomplete.

Aim and Methodology

The work described in this thesis was aimed at developing computational and statistical methods to guide the search and characterization of novel *cis*-regulatory elements. We pursued this through the integrated analysis of DNA sequence, gene expression, and epigenetic data.

Results

First, we analyzed the evolutionary history and species-specific divergence of *cis*-regulatory elements. Mutations in *cis*-regulatory elements have played a major role in species adaptation and speciation. Disruption of *cis*-regulatory elements has been associated with a wide range of human diseases. We found evidence arguing that TF binding site composition is often necessary to retain, and sufficient to predict regulatory activity in the absence of overt sequence conservation. Second, we addressed the sequence encryption of *cis*-regulatory elements and developed computational tools to decipher it. For this purpose, we collected distal *cis*-regulatory elements in the loci of genes expressed in particular cells and tissues, and constructed several machine learning classifiers that discriminate *cis*-regulatory elements from other noncoding sequences based on sequence features. Furthermore, we applied massive parallel testing to thousands of *de novo* designed *cis*-regulatory elements to evaluate and validate various regulatory grammar rules, including the effect on expression of the number of TF binding sites, their location, spacing, and order. Finally, we turned to the characterization of transcriptional networks partaking in embryonic development, in the search for suitable diagnostic and prognostic markers of congenital diseases.

Discussion and Conclusions

The aforementioned computational approaches and underlying mathematical models represent significant progress towards deciphering the genetic component of complex disease susceptibility. Our findings have been extensively validated with the aid of ChIP-seq and gene expression datasets, and, where applicable, compared to those produced by alternative methods. More importantly, collaborators from several experimental biology laboratories have independently confirmed our computational predictions in transgenic zebrafish and mouse. Transgenic experiments allow us to investigate transcriptional regulation in specific cells and tissues, as well as across embryonic development, studying both the enhancement and repression of transcription, using different model organisms. Eventually, the research line presented here promises to advance and complement the drug development efforts

to manage and prevent complex diseases such as diabetes and cancer, among other leading causes of death in the world.

Acknowledgements

I am using this opportunity to express my gratitude to everyone who supported me at the University of Rostock in Germany and at the National Institutes of Health in the United States of America in the past few years. I am thankful for their guidance, constructive criticism, and advice. In particular, I am deeply indebted to Prof. Dr. Georg Fuellen from the University of Rostock, for his encouragement and help to write this thesis, and to Dr. Ivan Ovcharenko from the National Institutes of Health, whose expertise and insight greatly contributed to my postdoctoral training.

A special acknowledgement is due to my family and friends for their endless love and patience. This thesis is dedicated to them.

"El motor de la ciencia es la curiosidad con las preguntas constantes: ¿Y eso cómo es? ¿En qué consiste? ¿Cómo funciona? Y lo más fascinante es que cada respuesta trae consigo nuevas preguntas. En eso los científicos le llevamos ventajas a los exploradores, cuando creemos haber llegado a la meta anhelada, nos damos cuenta de que lo más interesante es que hemos planteado nuevos problemas para explorar."

César Milstein (1927-2002)

Contents

List of Figures

1. Publications that are Part of this Thesis

1. **Taher L**, Smith RP, Kim MJ, Ahituv N, Ovcharenko I. Sequence signatures extracted from proximal promoters can be used to predict distal enhancers. *Genome Biol*. 2013;14(10):R117. PMID: 24156763.

2. Smith RP*, **Taher L***, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet*. 2013 Sep;45(9):1021-8. doi: 10.1038/ng.2713. Epub 2013 Jul 28. PMID: 23892608.

3. Visel A, **Taher L**, Girgis H, May D, Golonzhka O, Hoch RV, McKinsey GL, Pattabiraman K, Silberberg SN, Blow MJ, Hansen DV, Nord AS, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V, Kaplan T, Kriegstein AR, Rubin EM, Ovcharenko I, Pennacchio LA, Rubenstein JL. A high-resolution enhancer atlas of the developing telencephalon. *Cell*. 2013 Feb 14;152(4):895-908. doi: 10.1016/j.cell.2012.12.041. Epub 2013 Jan 31. PMID: 23375746.

4. Burzynski GM*, Reed X*, **Taher L***, Stine ZE, Matsui T, Ovcharenko I, McCallion AS. Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control. *Genome Res*. 2012 Nov;22(11):2278-89. doi: 10.1101/gr.139717.112. Epub 2012 Jul 3. PMID: 22759862.

5. Busser BW*, **Taher L***, Kim Y, Tansey T, Bloom MJ, Ovcharenko I, Michelson AM. A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLOS Genet*. 2012;8(3):e1002531. doi: 10.1371/journal.pgen.1002531. Epub 2012 Mar 8. PMID: 22412381.

6. **Taher L***, Narlikar L*, Ovcharenko I. CLARE: Cracking the LAnguage of Regulatory Elements. *Bioinformatics*. 2012 Feb 15;28(4):581-3. doi: 10.1093/bioinformatics/btr704. Epub 2011 Dec 22. PMID: 22199387.

7. **Taher L**, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I. Genome-wide identification of conserved regulatory activity in diverged sequences. *Genome Res*. 2011 Jul;21(7):1139-49. doi: 10.1101/gr.119016.110. Epub 2011 May 31. PMID: 21628450.

8. **Taher L**, Collette NM, Murugesh D, Maxwell E, Ovcharenko I, Loots GG. Global gene expression analysis of murine limb development. *PLOS ONE*. 2011;6(12):e28358. doi: 10.1371/journal.pone.0028358. Epub 2011 Dec 9. PMID: 22174793.

9. **Taher L**, Ovcharenko I. Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics*. 2009 Mar 1;25(5):578-84. doi: 10.1093/bioinformatics/btp043. Epub 2009 Jan 25. PMID: 19168912.

*These authors should be considered joint first authors.

2. Introduction

The human genome contains $3.2x10^9$ base pairs and an estimated of $23,000$ protein-coding genes (International Human Genome Sequencing Consortium 2004; Pennisi 2012). Furthermore, the estimated number of protein-coding genes in the human genome, has been repeatedly revised down from initial predictions of $100,000$ in the 1980s as genome sequence quality and gene finding methods have improved. By comparison, the mouse genome comprises $2.7x10^9$ base pairs and approximately the same number of protein-coding genes as the human genome (Mouse Genome Sequencing Consortium, Waterston, et al. 2002). Many plant genomes, such as maize, are relatively large and encompass more protein-coding genes than the human genome (Schnable, Ware, et al. 2009). There is no clear correspondence between the size of eukaryotic genomes and the number of protein-coding genes. Moreover, it is currently accepted that the number of genes bears no direct relationship to organismal complexity. The non-protein coding (noncoding) portion of the eukaryotic genome – $\sim 98.5\%$ in the case of humans – is associated with noncoding RNA or constitutes regulatory, structural and/or repetitive DNA (Lander, Linton, et al. 2001). Emerging evidence suggests that organismal complexity arises through the regulation of gene expression, and in particular, transcription (Levine and Tjian 2003).

2.1. Eukaryotic Transcriptional Regulation

In Eukaryotes, transcriptional regulation is achieved through the interaction of several levels of control, including chromatin packing and transcription factor (TF) activity.

Chromatin packing restricts the accessibility of the genes to the RNA polymerase and TF proteins. Indeed, chromatin exists in what is probably a continuum between the compact heterochromatin, tightly wrapped around the histones, and the open euchromatin. High-level transcription requires an open chromatin structure (Kornberg and Lorch 1992). Changes in chromatin structure, commonly referred to as chromatin remodeling, are regulated by the acetylation, methylation, and ubiquitination of histone tails (Strahl and Allis 2000).

TFs, also known as *trans*-acting elements, are regulatory proteins that can promote or inhibit transcription by binding to specific DNA sequences or *cis*-regulatory elements. All characterized transcribed protein-coding genes have a promoter immediately upstream of the 5' end of the transcription start site (TSS) that is recognized by the RNA polymerase II as well as basal TFs. Promoters consists of a core promoter, which is, in general, considered to be necessary and sufficient for low-level transcription, and additional proximal promoter elements (Smale and Kadonaga 2003). The core promoter is approximately 100 base pairs long, contains the TSS, and interacts directly with the components of the preinitiation complex (PIC) (Roeder 1996). Although the RNA polymerase II catalyzes RNA synthesis, it is unable to recognize the TSS or melt the DNA. Therefore, the RNA polymerase II cannot initiate transcription on its own. This is accomplished by the PIC, which involves general TFs, such as TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH (Lee and Young 2000). These TFs interact with a set of elements in the proximity of the TSS, such as the TATA box, the initiator element, the downstream promoter element (DPE), the TFIIB recognition element (BRE), and CpG islands. Eukaryotic core promoters are heterogeneous, and although most promoters contain one or more of these elements, none seems to be essential for promoter function (Smale and Kadonaga 2003). The core promoter is modulated by proximal promoter and additional distal *cis*-regulatory elements, which are recognized by more specific TFs. Distal *cis*-regulatory elements are similar to proximal promoters, but a single distal *cis*-regulatory element can control the expression of different genes at different times.

Distal *cis*-regulatory elements have been classified on the basis of their behavior in synthetic assays. Because the assays are well established, the distal *cis*-regulatory elements most studied to date are enhancers and silencers. Enhancers and silencers are typically a few hundred base pairs long, and mediate positive and negative regulation of transcription, respectively, independently of their position and orientation with respect to the TSS of their target gene. Thus, they may reside several thousands base pairs upstream or downstream of their target genes, even within introns of neighboring genes with unrelated expression patterns, and in some cases, in different chromosomes (Visel, Akiyama, et al. 2009; Williams, Spilianakis, et al. 2010). Although not well understood, distal transcriptional regulation would involve long-range direct interactions between the TFs bound to enhancers and silencers, and promoters, with concomitant looping of the intervening DNA. Furthermore, such interactions would be established either by simple diffusion within the nucleus or by an active “tracking” mechanism, in which enhancers and silencers migrate along the chromatin fiber until they encounter a promoter. TFs bound to enhancers, silencers, and promoters would interact with coactivators, proteins that recruit chromatin remodeling factors and communicate with the basal transcription machinery to assemble a functional PIC on the promoter (Bulger and Groudine 2011). A different class of distal *cis*-regulatory elements, known as insulators or boundary elements, establish discrete transcriptional domains. Insulators have been shown to possess either one or both of two activities, blocking enhancers (enhancer-blocking insulators) and/or protecting against heterochromatin spreading (barrier insulators) (Gaszner and Felsenfeld 2006). Most vertebrate enhancer-blocking insulators contain binding sites for the TF CTCF, a zinc finger protein with multiple roles, including transcription activation and repression (Bell, West, et al. 1999; Yusufzai and Felsenfeld 2004). CTCF has also been implicated in the function of barrier insulators, but it has been suggested that additional proteins may be required for specificity (Cuddapah, Jothi, et al. 2009). The above regulatory elements may be either discrete or clustered within locus control regions (LCRs) to mediate complex transcriptional programs involving several genes within a genomic locus.

2.2. *Cis*-regulatory Elements

Cis-regulatory elements consist of clusters of binding sites for TFs. Binding sites are short (usually 6-20 base pairs long) and degenerate sequences. The particular nature, number, and spatial arrangement of TF binding sites, together with the availability of the cognate TFs, determine the activity of a given *cis*-regulatory element, i.e., its effect on a gene. For example, the transcription of the *Drosophila Nidogen* (*Ndg*) gene at different developmental stages and in different cell types is controlled by the binding of multiple Forkhead (Fkh) TFs, which are differentially expressed in the developing *Drosophila*, each binding to distinct binding sites in the same enhancer (Philippakis, Busser, et al. 2006; Zhu, Ahmad, et al. 2012). In addition, we have recently shown that cardiac-specific expression of *Ndg* is dependent on the binding of Myb and a POU homeodomain (POUHD) TF (Figure 2.1). The number of binding sites for a given TF also plays a role in determining the activity of a *cis*-regulatory element. Indeed, more than 50% of known promoters and experimentally assayed enhancers in the vertebrate genome contain multiple binding sites for the same TF, making this a pervasive feature of vertebrate *cis*-regulatory elements. Furthermore, evolutionary conserved noncoding sequences containing multiple binding sites for the same TF occupy nearly 2% of the human genome, suggesting that arrangements of multiple binding sites for the same TF play an important role in transcriptional regulation, probably conferring robustness (Gotea, Visel, et al. 2010). The spacing between the TF binding sites also influences the interactions between the TFs that bind to a given *cis*-regulatory element. In order to engage in a direct protein-protein interaction, TF have to bind to sites positioned on the same face of the DNA. Since there are approximately 10 base pairs per helical turn of DNA, direct interactions between TFs are only possible if their binding sites are separated by a multiple of 10 base pairs. For example, the activity of a virus-inducible enhancer of the human *IFNB* gene requires a precise helical relationship between individual TF binding sites. Introducing slightly more than a half-helical turn (6 bp) between two known binding sites, reduces the level of virus induction by 9-fold. Inserting 10 bp fully restores the activity of the enhancer (Thanos and Maniatis 1995). Spacing rules between TF binding sites at larger distances usually reflect DNA wrapping around nucleosomes (Spitz and Furlong 2012).

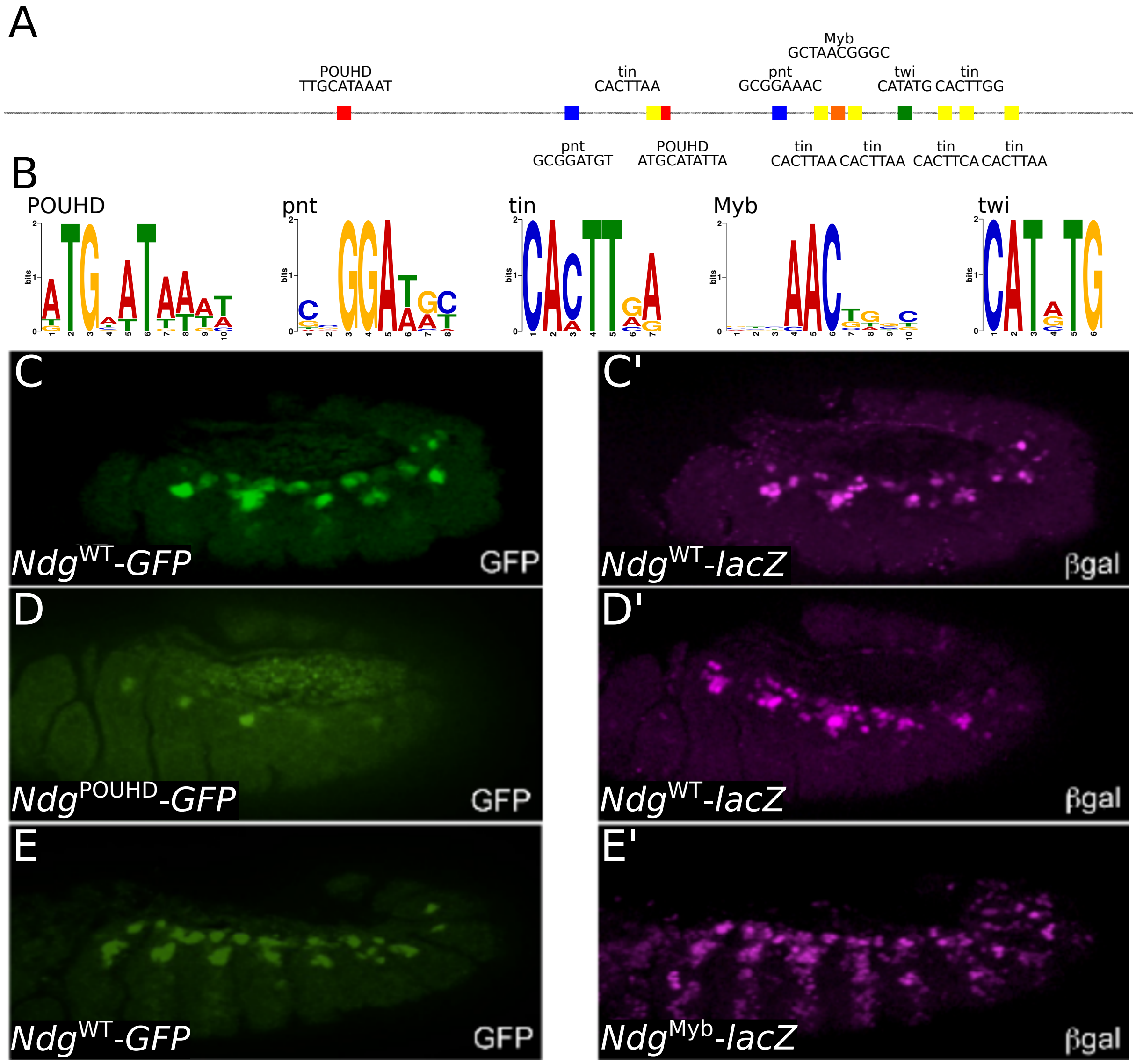

Figure 2.1. An enhancer of the *Drosophila melanogaster Ndg* gene. Wild-type regulatory activity of this enhancer requires the occurrence of binding sites for Myb and a POUHD TF. (A) Sequences representing binding sites for a POUHD TF, pointed (pnt), Myb, tinman (tin), and twist (twi) in the *Ndg* enhancer. (B) Motifs representing TF binding sites obtained from TRANSFAC (Matys, Kel-Margoulis, et al. 2006) and Philippakis, Busser, et al. 2006. (C) GFP (green) and the beta-D-galactosidase (*lacZ* (magenta) are co-expressed when driven by the wild-type (WT) *Ndg* enhancer (*Ndg*$^{\mathrm{WT}}$-*GFP* and *Ndg*$^{\mathrm{WT}}$-*lacZ*, respectively). (D) GFP (green) expression driven by a mutated *Ndg* enhancer, in which POUHD sites have been selectively inactivated (*Ndg*$^{\mathrm{POUHD}}$-*GFP*), is significantly reduced compared to beta-D-galactosidase (magenta) driven by *Ndg*$^{\mathrm{WT}}$-*lacZ*. (E) beta-D-galactosidase driven by a version of the *Ndg* enhancer in which Myb binding sites have been selectively inactivated (*Ndg*$^{\mathrm{Myb}}$-*lacZ*) is de-repressed into additional somatic mesodermal cells compared to GFP driven by a WT version of the *Ndg* enhancer (*Ndg*$^{\mathrm{WT}}$-*GFP*). Figure modified from Busser, Taher, et al. 2012.

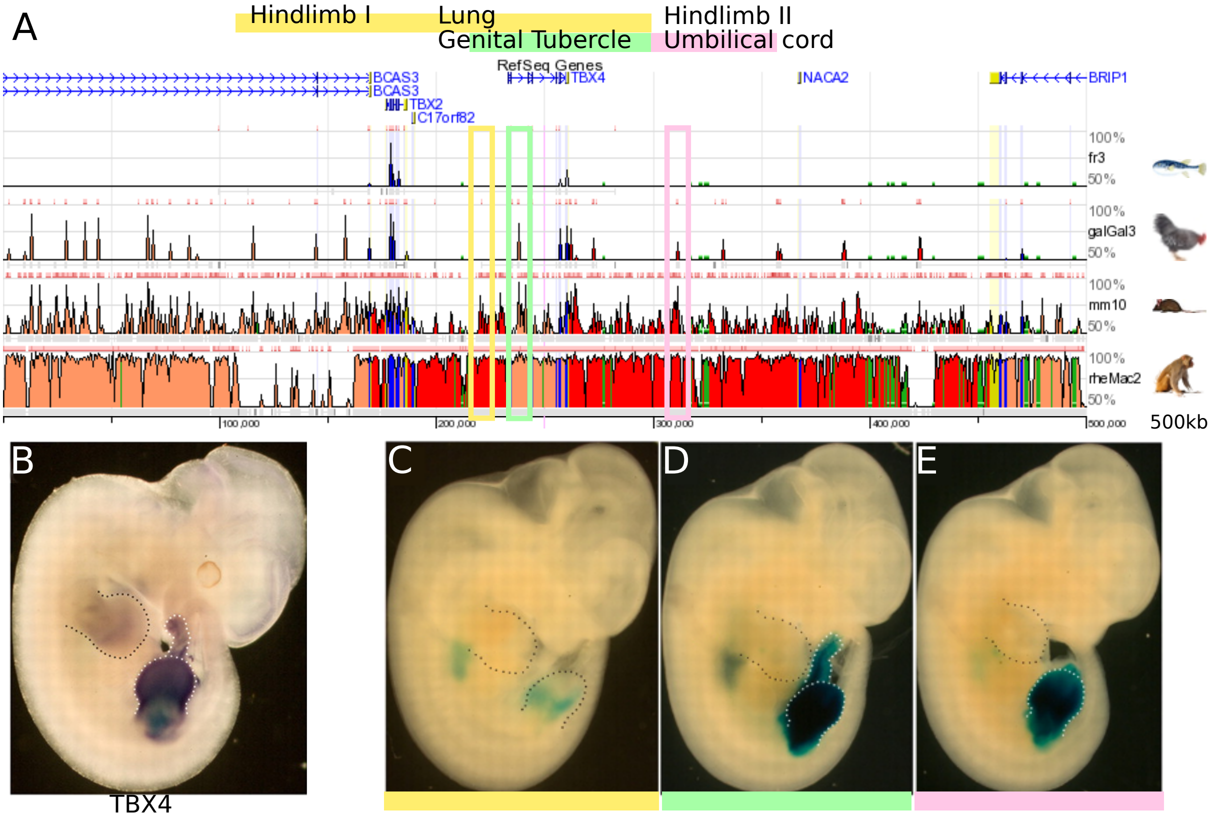

Figure 2.2. Region encompassing *TBX4* in the human genome (hg19). (A) Sequence comparison of human with chimpanzee, mouse, chicken and fugu (`http://ecrbrowser.dcode.org`). Regions with at least 70% identity over a 100 bp window are colored: blue, exons; red, conserved intergenic sequence; salmon, conserved intronic sequences; green, repeats. Colored boxes indicate positions of confirmed enhancers (see C, D, and E). (B) *In situ* hybridization for Tbx4 mRNA in E12.5 mouse embryos. (C-E) lacZ staining of E12.5 transgenic mouse embryos illustrating expression patterns seen with the different constructs: (C) yellow, hindlimb I; (D) green, lung and genital tubercle; (E) pink, hindlimb II and umbilical cord. Figure modified from Menke, Guenther, et al. 2008.

Spatiotemporal patterns of expression are largely controlled by distal *cis*-regulatory elements. Multiple distal *cis*-regulatory elements often act cooperatively or competitively (e.g, (Buttgereit 1993; Lin, Chen, et al. 2007; Perry, Boettiger, et al. 2011)) to determine the expression pattern of a gene. Thus, two independent enhancers control hindlimb expression in vertebrates, one located upstream and one downstream of the *TBX4* coding exons (Menke, Guenther, et al. 2008). These two enhancers differ in their precise patterns of activity within the hindlimb, and in their degree of sequence conservation (Figure 2.2). A third enhancer directs gene expression of *TBX4* in the lung. Decoupling transcriptional regulation into multiple *cis*-regulatory elements provides a flexible mechanism for altering the strength and location of the expression of a gene, making it possible, for example, to separately modify the size of forelimb and hindlimb bones during vertebrate evolution.

2.2.1. Using Reporter Gene Assays to Study *Cis*-regulatory Activity

The regulatory activity of genomic elements is routinary tested using reporter gene assays. Most assays use a relatively long DNA construct, such as a bacterial artificial chromosome (BAC), which encompasses the genomic region of interest and a reporter gene. The construct is introduced into a model system in order to observe and/or quantify the mRNA, protein or protein activity of the reporter gene after allowing time for gene expression. In a promoter assay, the putative promoter is placed directly in front of the reporter gene. In an enhancer assay, the putative enhancer is usually placed in front of a minimal promoter, which is not sufficient to direct gene expression without the presence of an enhancer, followed by a reporter gene. Insulators can be assayed for barrier insulator activity by placing them on both sides of a known enhancer and promoter, followed by a reporter gene.

An ideal reporter gene encodes a protein whose expression can be detected with high sensitivity above any endogenous expression and that displays a wide dynamic range of responses. Common reporter genes include those coding of the luciferase (*luc*), the green fluorescent protein (*GFP*), and the beta-D-galactosidase (*lacZ*).

2.2.2. High-throughput Identification of *Cis*-regulatory Elements

Chromatin immunoprecipitation used in conjunction with high-throughput sequencing (ChIP-seq) is one of the current methods of choice for the genome-wide interrogation of protein-DNA (e.g., TF-TF binding sites) interactions (e.g., (Mikkelsen, Ku, et al. 2007; Robertson, Hirst, et al. 2007)). Briefly, ChIP-seq assays involve (Landt, Marinov, et al. 2012):

(i) Cross-linking of proteins to the DNA, frequently perfomed by formaldehyde treatment.
(ii) Cell disruption and chromatin shearing by sonication or enzymatic digestion.
(iii) Immunoprecipitation with an antibody that is specific for the protein of interest (e.g., TFs, modified histone, RNA polymerase) with its bound DNA.
(iv) Reversion of the cross-linking.
(v) Analysis of the enriched DNA by high-throughput sequencing.

ChIP-chip is the array-based predecessor of ChIP-seq, where the DNA fragments of interest are hybridized on an array (Ren, Robert, et al. 2000).

Using genome-wide ChIP-chip, Heintzman, Hon, et al. (2009) characterized *cis*-regulatory elements according to their histone modification patterns on diverse human cell lines. Their analysis revealed that while the chromatin state at promoters and insulators is largely invariant across cell types, the chromatin state at enhancers is highly cell-specific, consistent with cell-specific activity. By performing ChIP-seq with antibodies against Ep300, a well-known coactivator, Visel, Blow, et al. (2009) were able to identify enhancers that are specifically active in the forebrain, midbrain and hindbrain of the developing mouse embryo with a 5-16 times higher success rate than that obtained based on sequence conservation alone. Similar strategies have been applied to other tissues and developmental stages in order to identify both tissue-specific and temporally important enhancers (May, Blow, et al. 2012). To date, multiple ChIP-seq maps of TF binding and histone modifications exist, constituting important resources for the investigation of the mechanisms involved in transcriptional regulation. In particular, the international Encyclopedia of DNA Elements (ENCODE) project, launched in 2003 by the National Human Genome Research Institute (NHGRI) with the aim of providing a global view of the functional elements encoded in the human genome sequence (ENCODE Project Consortium 2004), continues to generate vast amounts of ChIP-seq data across multiple tissues and cell types. These data are currently released under a rapid release policy, immediately after validation. Furthermore, the Mouse ENCODE project was initiated by the NHGRI in 2009 as a complement to the ENCODE project, in order to annotate functional elements encoded in the mouse genome by applying the same technologies and experimental pipelines developed for human ENCODE (Mouse ENCODE Consortium, Stamatoyannopoulos, et al. 2012).

Since transcriptionally active genomic regions are enriched with sites that are hypersensitive to DNases (Gross and Garrard 1988), high-throughput approaches targeting DNaseI hypersensitive sites (DHS) have proved useful for systematically uncovering *cis*-regulatory elements on a genome-wide scale. Briefly, cell nuclei are first digested with DNaseI. Cleaved DNA ends are then selected by different procedures, for example, by ligating them to a biotinylated tag which is then captured by a streptavidin column (Crawford, Holt, et al. 2004). Finally, the DNA encompassing DHSs is identified using microarrays or high-throughput sequencing. DNase-seq has been extensively used by the ENCODE consortium, which thereby determined the existence of millions of distinct DHS across the human genome, most of them presumably representing highly cell-specific distal *cis*-regulatory elements (Thurman, Rynes, et al. 2012).

ChIP-seq and DNase-seq are powerful techniques in that they can query the entire genome in a single experiment. However, these experiments identify *cis*-regulatory elements indirectly, based on their association with specific TFs, coactivators, histone modifications, and chromatin structure. That is, despite their relatively high success rates, these approaches are not functional assays. Therefore, they do not demonstrate by themselves the functional significance of a TF or modified histone associated with a particular genomic region. Indeed, putative *cis*-regulatory elements determined by ChIP-seq and/or DNase-seq must be validated with standard reporter gene assays (Thurman, Rynes, et al. 2012; Visel, Blow, et al. 2009).

Several methods have recently been developed or adapted to assess *cis*-regulatory activity with higher throughput than the standard reporter assays. For example, self-transcribing active regulatory region sequencing (STARR-seq) assays enhancer activity in a direct, quantitative, and genome-wide manner (Arnold, Gerlach, et al. 2013). STARR-seq works with any source of input DNA. Thus, Arnold, Gerlach, et al. (2013) sheared *Drosophila* genomic DNA and cloned the resulting fragments downstream of a minimal promoter. These constructs were then transfected into a *Drosophila* cell line. Active enhancers are assumed to enhance their own transcription. Hence, their activity can be estimated from their level of transcription. Unlike STARR-seq, site-specific integration fluorescence-activated cell sorting followed by sequencing (SIF-seq, (Dickel, Zhu, et al. 2014)) seeks the integration of a putative *cis*-regulatory element into a single genomic locus, providing a reproducible chromosomal context. Putative *cis*-regulatory elements are linked to a minimal promoter and a gene encoding the Venus yellow fluorescent protein, and targeted into a single site in the genome of a mouse embryonic stem cell. Cells expressing the reporter gene are isolated using flow cytometry, and the *cis*-regulatory elements inserted into these cells are finally amplified and identified through high-throughput sequencing. The use and development of high-throughput methods able to assess regulatory activity in a genomic context and in a wide variety of cells will eventually enable the comprehensive study of the roles of distal *cis*-regulatory elements in the human genome.

2.3. Prediction of *Cis*-regulatory Elements

The variability in the genomic location and sequence heterogeneity of distal *cis*-regulatory elements makes them particularly difficult to predict (Hardison and Taylor 2012).

Since functionally relevant sequences are under purifying selection, they are generally more conserved than non-functional sequences. On these grounds, comparative genomics rapidly became one of the most widely used strategies to predict distal *cis*-regulatory elements. One of the first applications of comparative genomics to the identification of distal *cis*-regulatory elements involved a 1,000,000-base-pair-long human locus comprising the *IL4*, *IL13*, and *IL5*, three cytokine genes, which had been previously shown to exhibit coordinated expression (Kelly and Locksley 2000). This locus also encompasses approximately 90 highly conserved noncoding sequences (CNSs, at least 70% sequence identity with mouse over 100 base pairs). The largest CNS was assayed for regulatory activity in transgenic and knockout mouse assays, concluding that this element modulates the expression of all three cytokine genes, separated by more than 120,000 base pairs of sequence (Loots, Locksley, et al. 2000). More generally, comparative genomics has proven to be powerful, in that, ∼ 50% of highly conserved noncoding sequences act as enhancers in reporter gene assays (Bejerano, Pheasant, et al. 2004; Pennacchio, Ahituv, et al. 2006). Interestingly, most assayed CNSs with regulatory activity (available through the VISTA Enhancer Browser, `http://enhancer.lbl.gov`, (Visel, Minovitsky, et al. 2007)) direct gene expression in the central nervous system.

Computational methods attempting to identify distal *cis*-regulatory elements in the vertebrate genome face several challenges:

(i) The modeling and prediction of TF binding sites is usually done using position-weight matrices (PWMs, (Staden 1984)) collected from the literature. PWMs available in motif databases describe the probability of observing the respective nucleotides A, C, G, and T in each position of a sequence. Examples of collections of PWMs include TRANSFAC (Matys, Kel-Margoulis, et al. 2006) and JASPAR (Sandelin, Alkema, et al. 2004). However, TF binding sites are normally short and highly variable, and thus, functionally significant binding events can only be predicted with relatively low sensitivity and specificity.

(ii) Genome-wide analysis of ENCODE ChIP-seq datasets shows that co-binding TF often exhibit position and orientation preferences (Jankowski, Szczurek, et al. 2013; Wang, Zhuang, et al. 2012). However, it remains to be learned to what extent constrained spacing and orientation of interacting TFs are critical for regulatory activity.

(iii) Transcriptional regulation usually requires the coordinated action of multiple TFs. Indeed, the combinatorial binding of TFs to the DNA defines the spatial and temporal expression patterns driven by a given *cis*-regulatory element. Most of our knowledge regarding synergistic and antagonistic interactions between TFs has been derived from *in vitro* experiments. However, evidence is accumulating that the epigenome can modulate TF cooperativity (Chen, Xiao, et al. 2013).

(iv) Additional, possibly unknown, genetic and epigenetic factors may be important for *cis*-regulatory activity. For example, enhancers active in the heart have a significantly higher GC content than enhancers active in other tissues (Erwin, Oksenberg, et al. 2014).

Limited by our current understanding of transcriptional regulation, most computational methods for predicting distal *cis*-regulatory elements are based on general genomic features: the nature and number of TF binding sites, their spatial constraints (e.g., putative *cis*-regulatory element have high densities of TF binding sites), and/or their evolutionary conservation. Existing methods can be roughly classified into four families:

(i) Evolutionary methods that compare homologous sequences between distant or closely related species (e.g., (Boffelli, McAuliffe, et al. 2003; Ovcharenko, Stubbs, et al. 2004)).

(ii) Probabilistic models that search for significant clusters of TF binding sites (e.g., (Blanchette, Bataille, et al. 2006; Frith, Li, et al. 2003; Narlikar, Sakabe, et al. 2010)).

(iii) Machine learning approaches that scan the genomic sequence for windows containing multiple motif matches to known TF binding sites (e.g., (Busser, Taher, et al. 2012)).

(iv) Alignment-like methods, that align the sequence (i.e., order) of TF binding sites found on a sequence to that of a known *cis*-regulatory element (e.g., (Hallikas, Palin, et al. 2006)).

Many methods are hybrids of two or more strategies (Su, Teichmann, et al. 2010).

2.4. *Cis*-regulatory Elements in Human Diseases

Most mutations in regulatory sequences do not disrupt the amino acid sequence of genes, do not create alternative transcripts, do not introduce premature stop codons, and do not affect the 3-dimensional structure of proteins. Instead, they affect the dynamics of gene expression. Noncoding mutations in *cis*-regulatory elements controlling TFs can radically alter entire regulatory networks, causing, for example, changes in cell-fate decisions. However, the overall expression pattern of a gene is usually controlled by multiple *cis*-regulatory elements. As a result, the majority of noncoding mutations are likely to have cell- and condition-specific, rather than pleiotropic effects (Carroll 2008). Thus, while mutations within an enhancer 1.5 million base pairs upstream of *SOX9* result in Pierre Robin sequence (PRS), a form of cleft palate, mutations within other two regulatory elements result in bone dysplasia and in disorders of sex development, respectively (Benko, Gordon, et al. 2011; Benko, Fantes, et al. 2009; Kurth, Klopocki, et al. 2009). On the other hand, coding mutations in *SOX9* lead to the campomelic dysplasia syndrome, which comprises PRS, skeletal defects and sex reversal (Wagner, Wirth, et al. 1994). Hence, regulatory mutations are currently assumed to constitute a significant determinant of disease (Epstein 2009; Herz, Hu, et al. 2014; Smith and Shilatifard 2014; Symmons and Spitz 2013).

To directly assess the impact of a noncoding mutation on gene expression, Genome-Wide Association Studies (GWAS) test for correlations between gene expression in unrelated individuals and SNP and copy-number variants (CNV) genome-wide profiles (Stranger, Nica, et al. 2007). GWAS have revealed an ever-expanding list of diseases associated with noncoding genetic variants (Manolio, Collins, et al. 2009). For example, complex diseases such as Alzheimer's, type 2 diabetes, multiple sclerosis, and cancer, all with high socio-economic impact, are associated with noncoding SNPs (Stranger, Stahl, et al. 2011). In turn, over 90% of the $\sim 15,000$ SNPs in the manually curated GWAS Catalog of the NHGRI (`www.genome.gov/gwastudies`, (Welter, MacArthur, et al. 2014)) reside in noncoding portions of the genome. Moreover, ENCODE has recently shown that 76% of noncoding SNPs are indeed in or very near DHSs, which suggests they are likely regulatory in nature.

Furthermore, a general model considers regulatory mutations as the key driving force behind the evolution of species, first advocated in a pioneering work of King and Wilson in 1975 (King and Wilson 1975). For instance, lactose tolerance has evolved recently, and varies among different human populations (characteristic to about 90% of Americans of northern European descent, 90% of Africa's Tutsi tribe, 50% of French, and 1% of Chinese, for example). Since lactase's only function is the digestion of lactose in milk, most mammals cease to produce lactase, which is coded for by the *LCT* gene, after weaning (Swallow 2003). In some European and African populations, the ability to digest milk through adulthood was independently fixed since cattle domestication in the early Neolithic Era (Tishkoff, Reed, et al. 2007). More precisely, the expression of *LCT* through adulthood depends on a regulatory mutation located $\sim 20,000$ base pairs upstream of *LCT* (Enattah, Sahi, et al. 2002; Wang, Harvey, et al. 1995).

Finally, over 30% of SNPs in the GWAS Catalog of the NHGRI are located at least 10,000 base pairs away from any known protein-coding gene, arguing for the importance of distant regulatory mutations in disease susceptibility and phenotypic diversity in the human population.

3. Novel Methods and Results

Biological sequence data are rapidly accumulating due to the progress of DNA sequencing technologies. Currently available genome-wide datasets have the potential to dramatically impact our understanding of the mechanisms that control gene expression, and, ultimately, of how gene expression relates to development, physiology and disease. We have developed a series of integrative approaches combining the development of novel computational methods, statistical analysis of genetic, epigenetic and expression data, and experimental validation using transgenic assays with the aim to determine how DNA sequence features and variations contribute to *cis*-regulatory activity.

3.1. Most Noncoding Sequences in the Human Genome are Associated with TF and Developmental Genes

A basic observation of genome biology is that genes differ widely in their size (length) and structure within and between species. Intergenic regions also vary. Furthermore, in the eukaryotic genome, the size of the intergenic regions is strongly correlated with the regulatory complexity of their corresponding flanking genes (Nelson, Hersh, et al. 2004). Hence, it is not surprising that genes associated with different biological functions have very different intergenic sizes (Taher and Ovcharenko 2009). This has important implications for the functional analysis of *cis*-regulatory elements.

The goal of functional enrichment analysis is to determine whether a set of coding or noncoding genomic elements is statistically enriched for some biological annotation. Many annotation databases have been developed, including the Gene Ontology (GO) (Ashburner, Ball, et al. 2000). A common test to calculate enrichment is Fisher's Exact Test. Let G be the set of all N genomic elements, k of which are annotated with a particular GO term. Let S be a subset of G containing n elements. The probability that m elements in G are annotated with that GO term can be calculated as:

$$P(X \geq m|n,k,N) = \sum_{i=m}^{\min(n,k)} \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}}$$

The standard approach to annotating *cis*-regulatory elements is to annotate them according to the one or two nearest genes. However, this procedure introduces a strong bias toward genes that are flanked by long intergenic regions (Taher and Ovcharenko 2009). For example, the majority of genes with relatively long intergenic regions relate to basic cellular processes, such as cell adhesion, binding, TFs, and development), while genes with relatively short intergenic regions correspond to lineage-specific and adaptive feature. Consequently, annotating *cis*-regulatory elements according to the one or two nearest genes leads to the false inference of over- and under-representation of specific GO categories that preferentially contain longer or shorter genes, respectively. As an alternative, we proposed to use a binomial distribution with parameters n and p_{GO}. p_{GO} is the probability of observing an association with a particular GO term, assuming that we are randomly sampling noncoding elements from the human genome:

$$p_{GO} = \frac{L_{nc}^{GO}}{L_{nc}^{HG}}$$

where L_{nc}^{GO} is the total size of the noncoding sequence in the loci of genes annotated with a given GO term, and L_{nc}^{HG} is the total size of the noncoding sequence in the human genome. Given a set S containing n noncoding elements, the probability that m or more elements in S are annotated with the GO term can be calculated as:

$$P(X \geq m|n, p_{GO}) = 1 - \sum_{i=0}^{m-1} \binom{n}{i} \cdot p_{GO}^{i} \cdot (1 - p_{GO})^{n-i}$$

Accounting for the variability in the size of intergenic regions across different gene loci effectively eliminates the ascertainment bias from the functional characterization of noncoding elements.

3.2. Noncoding Sequence Conservation is not Necessary nor Sufficient for *Cis*-regulatory Activity

Although comparative genomics has proven effective in identifying known and novel *cis*-regulatory elements in the human genome, the degree of noncoding sequence conservation is not directly correlated with functional constraint. For example, targeted deletion of ultraconserved sequences in mice failed to reveal notable abnormalities (Ahituv, Zhu, et al. 2007). Also, a study relying on ChIP-seq experiments with the coactivator Ep300 indicated that *cis*-regulatory elements involved in heart development are only weakly conserved across vertebrates (Blow, McCulley, et al. 2010). Furthermore, only a small fraction of presumably functional TF binding sites appear to be shared between mammals (Kunarso, Chia, et al. 2010; Schmidt, Wilson, et al. 2010). Indeed, the prevailing consensus is that the function of *cis*-regulatory elements is maintained by the conservation of their overall architecture, rather than of the sequence.

To recognize *cis*-regulatory elements that share a common ancestor but have diverged to the point where we cannot align them using standard comparative genomic approaches, we designed a strategy that relies on pairwise alignments among at least three species. Specifically, given $N > 2$ species, we aim at identifying triplets of sequences for which we reliably align (e.g., with at least 70% sequence identity over 100 base pairs) at most $\frac{N \cdot (N-1)}{2} - 1$, and not less than 2 out of the $\binom{N}{2}$ possible pairs. For example, the human and zebrafish sequences in Figure 3.1 cannot be confidently aligned. However, the alignments between zebrafish and frog, and frog and human have a sequence identity greater or equal than 70%, suggesting that all three sequences most likely share a common ancestor. Moreover, since the human sequence is relatively well conserved, approximately half of these sequences are expected to be functional (Pennacchio, Ahituv, et al. 2006), and constitute diverged *cis*-regulatory elements. We systematically scanned the human, frog, and zebrafish genomes and identified $\sim 1,500$ analogous triplets of sequences. These sequences were next employed to construct a model to describe pairs of diverged orthologous *cis*-regulatory elements that preserve their ancestral function (Taher, McGaughey, et al. 2011).

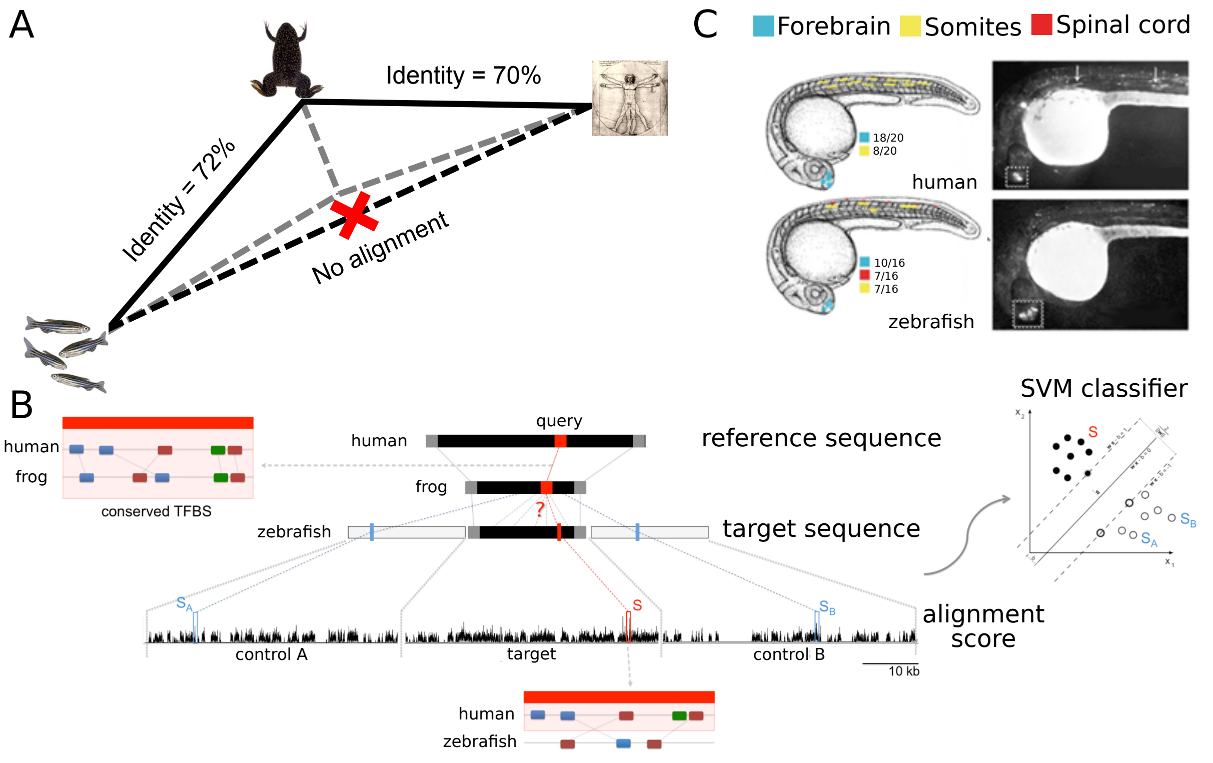

Figure 3.1. Alignment-free prediction of orthologous *cis*-regulatory elements. (A) Pairwise alignments for three orthologous sequences in human (hg18, chr18:53,271,349-53,271,555), frog (xenTro2, scaffold_97:133,388133,595), and zebrafish (danRer5, chr24:28,243,171-28,243,307). Only the human and the frog sequences and the frog and the zebrafish sequences can be aligned with at least 70% identity across at least 100 base pairs. The frog sequence has evolved more slowly relative to the human and zebrafish sequences, and thus, can be used to establish the orthology of the diverged human and the zebrafish sequences. (B) Overview of the detection of diverged orthologous *cis*-regulatory elements. We look for functional orthologs of conserved human/frog CNSs (*query*) in the zebrafish sequence by computing TF binding site alignments for the *target* and *control* loci, and using a support vector machine (SVM) to distinguish significant from random alignments. (C) Composite overviews of *in vivo* GFP expression data from 16-20 individual zebrafish embryos for constructs encompassing a candidate human (hg18, chr1:7,633,413-7,633,621) and zebrafish (danRer5, chr23:28,890,355-28,890,563) pair. The keys for the marked expression are provided next to each image, followed by the number of fish in the set with that specific expression. One representative GFP live image from each enhancer set is displayed. Figure modified from Taher, McGaughey, et al. 2011.

Since the functional units of *cis*-regulatory elements are TF binding sites, our model relies on the binding site composition of the sequences, rather than on their nucleotide composition. Let us assume a *cis*-regulatory element that is conserved according to a number of criteria between two species A and B and not in C. To identify its diverged functional ortholog(s) in C, we:

(i) Delimit syntenic loci in A and B encompassing the *cis*-regulatory element.
(ii) Delimit a *target* locus in C that is orthologous to the loci delimited in (i).
(iii) Define two *control* loci in C, adjacent to the *target* locus and with the same length.
(iv) Search and score putative TF binding sites in the orthologous *cis*-regulatory elements in A and B using the motif databases TRANSFAC (Matys, Kel-Margoulis, et al. 2006) and JASPAR (Sandelin, Alkema, et al. 2004) and MAST (Bailey and Gribskov 1998).
(v) Search and score putative TF binding sites in the locus delimited in C, analogously as in (iv).
(vi) Score the sequence of the *target* and *control* loci in C using a sliding window approach, looking for sequences of putative TF binding sites that are observed in both A and B.
(vii) Compare the scores of the best-scoring window in the *target* and *control* loci using a support vector machine (SVM), to decide whether the best-scoring window in the *target* locus represents an ortholog of the *cis*-regulatory element in A and B.

A model trained on the dataset of $\sim 1,500$ human, frog, and zebrafish sequences mentioned above was able to predict the correct zebrafish ortholog for 51% of the human sequences. The model was next applied to a more general case, where we cannot trace the ancestry of the human and zebrafish sequences using the frog sequence. We found candidate pairs of human/zebrafish diverged orthologous *cis*-regulatory elements in 10% of the loci that we examined.

The human counterparts of the candidate pairs are conserved in frog, enriched in binding events for the coactivator Ep300, and in the neighborhood or genes associated with development and TF activity, suggesting that they act as *cis*-regulatory elements *in vivo*. Transgenic zebrafish assays were undertaken to verify the function of a set of randomly selected 18 candidate pairs. As expected based on the level of conservation of the sequences, 8/18 (44%) of the assayed human sequences displayed enhancer activity. More importantly, 7/8 (88%) of the assayed zebrafish sequences also exhibited enhancer activity. Moreover, 5/7(71%) of the human/zebrafish pairs showed consistent patterns of activity, confirming that our method is able to identify orthologous *cis*-regulatory element despite extensive sequence divergence.

3.3. Sequence Features Predict *Cis*-regulatory Activity

Standard comparative genomic approaches to predict *cis*-regulatory elements rely on the assumption that functional elements are conserved across species. However, as illustrated in the previous section, this assumption is not always justified. In addition, the power of comparative genomics is limited in that it cannot determine the saptiotemporal pattern of expression driven by a particular *cis*-regulatory element. In order to overcome this restriction, we designed an approach that discriminates *cis*-regulatory elements from other noncoding sequences in the genome using sequence motifs as features. Since *cis*-regulatory elements basically consist of clusters of TF binding sites (Taher, Narlikar, et al. 2015), we represent each *cis*-regulatory element by a vector of binding site occurrences, analogously to the spectrum kernel (Leslie, Eskin, et al. 2002). Specifically, given a *cis*-regulatory element denoted by S_i and a set of n motifs corresponding to TF binding sites denoted by M_1, M_2, ..., M_n, the feature vector representation of S_i is:

$$v = \begin{bmatrix} f_{i1} \\ f_{i2} \\ \cdots \\ f_{in} \end{bmatrix}$$

where f_{ij} denotes the (frequency of) occurrence of motif M_j in sequence S_i.

The approach involves the following steps:

(i) Compile a collection of *cis*-regulatory elements with a particular activity (e.g., drive expression in the developing hindbrain) (*positive set*).

(ii) Randomly sample genomic regions of the same length and sequence properties (e.g., GC- and repeat-content) as the elements in the *positive set* (*control set*).

(iii) Search and score putative TF binding sites in the sequences of the elements compiled in (i) and (ii), using:

  a) *known* motifs from databases such as TRANSFAC (Matys, Kel-Margoulis, et al. 2006) and JASPAR (Sandelin, Alkema, et al. 2004); and/or

  b) *de novo* motifs, discovered by looking for significantly overrepresented sequence patterns within the sequences of interest with tools such as MEME (Bailey and Elkan 1994) and PRIORITY (Narlikar, Gordân, et al. 2006).

(iv) Represent each element in (i) and (ii) by a fixed-dimension feature vector of (frequency of) occurrence of the motifs in (iii).

(v) Construct a machine learning classifier to distinguish the feature vectors representing elements in the *positive set* from the feature vectors representing elements in the *control set*.

We used different machine learning algorithms to construct models that describe sets of enhancers that drive gene expression in particular contexts, such as the developing vertebrate hindbrain (Burzynski, Reed, et al. 2012) and forebrain (Pattabiraman, Golonzhka, et al. 2014; Visel, Taher, et al. 2013), and mesoderm and somatic muscle in *Drosophila* (Busser, Taher, et al. 2012).

For instance, in order to determine the sequence basis of transcriptional regulation in the vertebrate hindbrain (Burzynski, Reed, et al. 2012), we first compiled a dataset of 211 sequences from the human genome that had been shown to act as enhancers in the mouse and/or zebrafish hindbrain in *in vivo* reporter assays. Next, for each enhancer we randomly selected 10 *control* sequences with similar length, GC, and repeat-content from the noncoding portion of the genome. Then, we trained three SVM classifiers on different but highly overlapping subsets of the dataset to distinguish enhancers from their respective *control* sequences. To evaluate the classification performance we measured the area under the Receiver operating characteristic (ROC) curve (AUC) in a cross-validation setup. The SVM classifiers achieved similar performances, with an average AUC of $\sim 90\%$, indicating that they are able to accurately discriminate hindbrain enhancers from other noncoding sequence in the genome (Figure 3.2A). Finally, we applied the classifiers to predict novel hindbrain enhancers in the human genome, and identified a total of $\sim 40,000$ sequences that were classified as hindbrain enhancers by all classifiers, constituting good hindbrain enhancer candidates. A random subset of 34 candidates were tested using transgenic zebrafish assays (Figure 3.2B). In 88% of the cases, the results of the assays verified strong hindbrain enhancer activity (Figure 3.2C). In contrast, none of the 6 sequences selected among deeply conserved sequences (determined using the Most Conserved Elements database from the UCSC Table Browser, (Siepel, Bejerano, et al. 2005)) that were not classified as hindbrain enhancers by any of the classifiers drove consistent expression in the hindbrain. Notably, our validation rates are similar to those obtained using ChIP-seq with Ep300 experiments (Visel, Blow, et al. 2009), confirming the high predictive power of our computational model (Figure 3.2D).

In addition, our computational models are able to distinguish between enhancers that are active in different parts of a tissue, such as the mammalian embryonic forebrain (Visel, Taher, et al. 2013). Thus, to improve our understanding of the *cis*-regulatory architecture and gene networks relevant to forebrain development, we combined sequence conservation and ChIP-seq experiments. Thereby, we identified a total of $4,656$ candidate embryonic forebrain enhancers. 105 of 329 (32%) of these candidates tested in *lacZ* reporter assays in transgenic mice exhibited reproducible forebrain enhancer activity at embryonic day (E)11.5. The precise spatial expression pattern of the 105 enhancers was subsequently determined through a high-resolution analysis. We found that $\sim 40\%$ of the enhancers were active in the dorsal part of the embryonic forebrain, or pallium, while another $\sim 40\%$ were active in the ventral part of the embryonic forebrain, or subpallium. The remaining

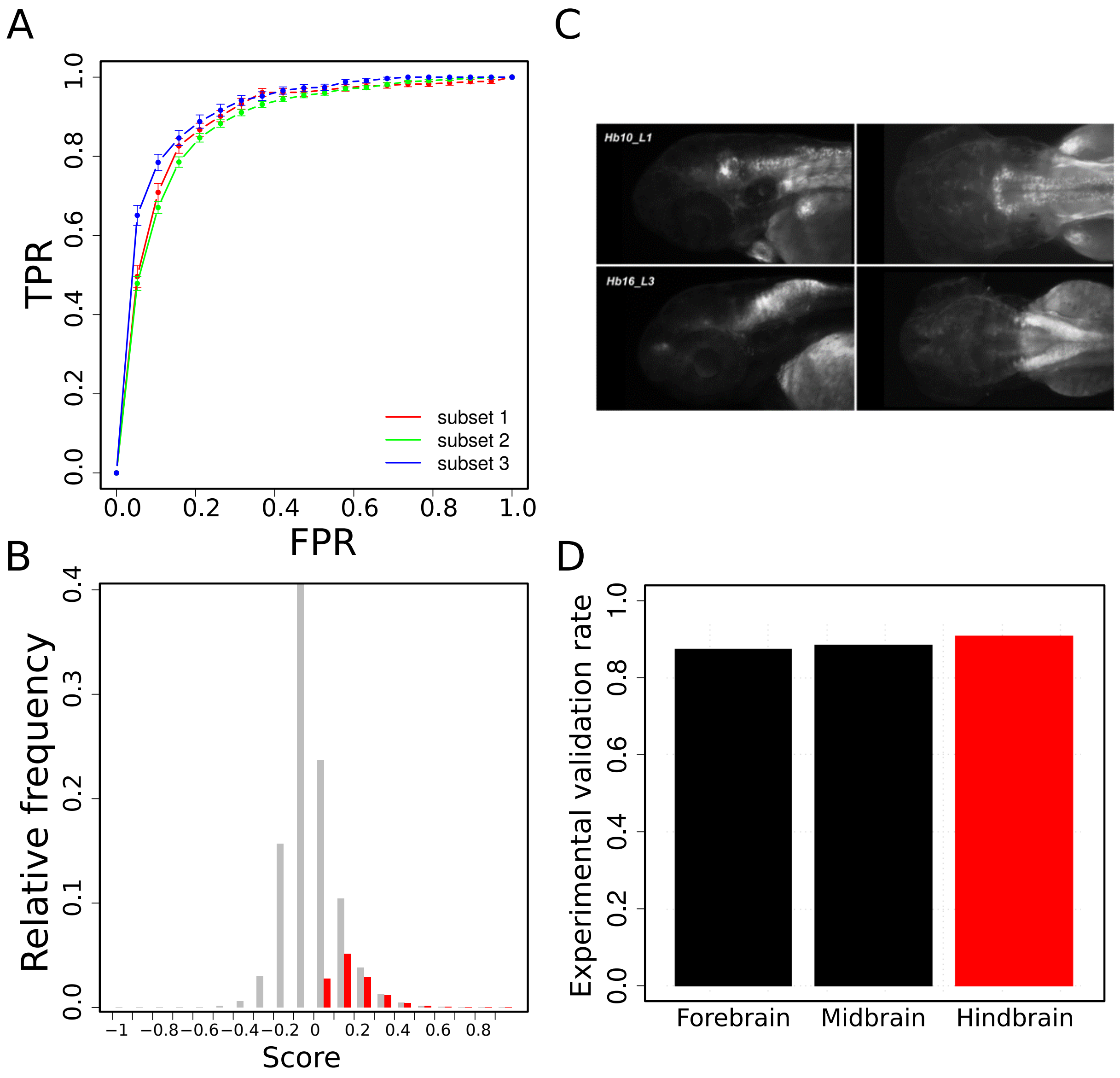

Figure 3.2. Systematic elucidation and *in vivo* validation of enhancers active in the vertebrate hindbrain. (A) Area under the ROC curve (AUC) for three hindbrain enhancer support vector machine (SVM) classifiers trained on three subsets of the complete dataset. AUC values range from 0.5 (random discrimination) to a theoretical maximum of 1. The average AUC for the three hindbrain classifiers is ~ 0.9. (B) Combine SVM score distribution for noncoding sequences in the human genome, for all three classifiers (gray). Scores have been transformed into $[-1, 1]$, preserving the sign. Scores > 0 correspond to candidate hindbrain enhancers for at least one classifier. Approximately 12% of the sequences had a score > 0 for all three classifiers (red). (C) GFP reporter expression from stable lines corresponding to hindbrain candidate enhancers showing expression across the hindbrain as well as in some non-hindbrain domains. (D) Validation rates in *in vivo* reporter assays for forebrain and midbrain candidate enhancers determined using ChIP-seq with Ep300 experiments (Visel, Blow, et al. 2009) as compared to the hindbrain candidate enhancers predicted using our SVM classifiers.

$\sim$ 20% were active in both pallium and subpallium. We constructed a Random Forest (RF) classifier to discriminate between these three classes of enhancers based on motif occurrences. In contrast to SVMs, which are inherently two-class classifiers, RF classifiers naturally support multiclass classification. Furthermore, RFs can be used to directly rank the importance of variables by comparing the error rate in the classification with the error rate obtained when the values of a variable are randomly permuted (Breiman 2001). Our RF classifier accurately predicted the activity of $\sim$ 80% of the enhancers. Moreover, the motifs that were found to be highly important for the classification were overall more evolutionary conserved than non-important motifs, supporting their functional relevance. Indeed, the majority of the most discriminatory motifs corresponded to predicted binding sites for homeodomain-containing TFs, consistent with the known critical role of these proteins in forebrain development (Hébert and Fishell 2008). In particular, we found motifs compatible with the binding sites of several members of the distal-less homeobox (DLX) TF family. This example demonstrates how various computational and experimental tools can be combined to investigate gene regulatory mechanisms underlying embryonic development, and ultimately understand the role of distal *cis*-regulatory elements in developmental disorders.

3.3.1. *Cis*-regulatory Sequence Features are Conserved in Diverged Orthologs

We next combined SVMs with comparative genomics to further elucidate the regulatory mechanisms underlying the formation of mesoderm and muscle founder cells (FCs) in *Drosophila melanogaster* (Busser, Taher, et al. 2012). The *Drosophila* mesoderm gives rise to a number of tissues, including heart, fat body, and visceral and somatic muscle. Among other TFs, twist has been shown to be essential for mesoderm development (Baylies and Bate 1996), and, in particular, in the specification of muscle types (Furlong, Andersen, et al. 2001). A defined number of cells in the mesoderm segregate as muscle progenitor cells, which further divide to become muscle FCs (Taylor 2000). Each somatic muscle derives from the fusion of a single muscle FC with a set number of fusion-competent cells.

Because the dataset of muscle FC enhancers that was available to us comprised only 16 sequences, we first devised a phylogenetic profiling strategy to extract diverged orthologs (50-80% sequence identity) of each *Drosophila melanogaster* muscle FC enhancer in 14 other insects, including 11 *Drosophila* species (Siepel, Bejerano, et al. 2005). At most two orthologs were selected for each *Drosophila melanogaster* muscle FC enhancer, totaling 24 sequences from 6 different species. Transgenic reporter assays confirmed that, despite extensive evolutionary reshuffling of known critical TF binding sites, the orthologs directed gene expression in patterns comparable to those of their *Drosophila melanogaster* counterparts. We also showed that including the enhancer orthologs in the motif enrichment analysis increased our statistical power to detect subtle patterns and associations in the original data.

The dataset comprising the 16 *Drosophila melanogaster* muscle FC enhancers and their 24 orthologs was used to train a SVM classifier to distinguish muscle FC enhancers from other noncoding sequences in the *Drosophila melanogaster* genome. We assessed the performance of the classifier in a cross-validation setup, obtaining an average AUC of $\sim$ 0.89, which demonstrates the accuracy of our approach. We used this classifier to achieve two main aims:

(i) Create a genome-wide map of putative muscle FC enhancers.
(ii) Establish the identify of TFs with a relevant role in muscle FC transcription.

First, we predicted a total of 5,500 muscle FC enhancers, which were found to be 4-fold overrepresented in proximity to genes that are known to be expressed in muscle FCs. Moreover, *in situ* hybridization showed that many genes with unknown functions flanking muscle FC enhancer predictions are indeed expressed in muscle FCs (13-fold enrichment, $P = 0.0002$). Transgenic reporter assays validated 75% of our enhancer predictions. Second, we confirmed that many of the TFs associated with the TF binding sites exhibiting the greatest power in discriminating muscle FC enhancers from other noncoding sequences in the genome are involved in muscle development. We recognized

Myb, ETS, POU homeodomain (POUHD), forkhead, and T-box motifs as critical for muscle FC transcription, a role subsequently validated through reporter and mutagenesis assays. In conclusion, our results show that machine learning combined with comparative genomics is useful for recognizing functional TF binding sites and for facilitating the identification of cognate TFs that control specific spatiotemporal patterns of gene expression.

3.3.2. Proximal and Distal *Cis*-regulatory Elements Share Sequence Features

Finally, we developed a machine learning framework to show that proximal and distal *cis*-regulatory elements cooperate very closely to determine the spatiotemporal pattern of expression of a gene (Taher, Smith, et al. 2013). Since promoters are typically immediately adjacent to the TSS of their target genes, they can be predicted with relatively high accuracy (Narlikar 2014). Indeed, promoter prediction is a common element of gene prediction methods. In contrast, enhancers can be located virtually anywhere in the genome. However, it has been long recognized that promoters and enhancers are functionally very closely related, and possibly indistinguishable from each other (e.g., (Maniatis, Goodbourn, et al. 1987; Maston, Evans, et al. 2006)). All of this directed us to the assumption that we can use the features present in the promoter sequence to predict enhancers.

For this purpose, we first compiled 79 datasets of promoters based on gene expression profiles on different human tissue and cell types, which we derived from the GNF Novartis Gene Expression Atlas (Su, Wiltshire, et al. 2004). Each dataset consisted of the 200 promoters of the most highly (*positive set*) and most lowly expressed (*control set*) genes in a given tissue or cell. For each tissue or cell, predicted TF binding sites within the sequences of the *positive* and *control* sets were used to train a SVM classifier capable of distinguishing between the two sets. We obtained reliable classifiers for 92% (73/79) of the tissues and cells under consideration, with an AUC between 60% (for subthalamic nucleus promoters) and 98% (for heart promoters). Basically, this result provides evidence for the occurrence of features in the promoter sequence that are associated with the spatiotemporal pattern of expression of its target gene.

We next applied the classifiers to predict enhancers, scanning the noncoding sequence of the loci of the 200 most highly and lowly expressed genes in each of the 73 tissues with reliable classifiers. Thirty-percent of reliable promoter-based classifiers produced consistent enhancer predictions, with significantly higher densities in the loci of the most highly expressed compared to lowly expressed genes (e.g., over 5-fold enrichment in the case of liver). Enhancer predictions were verified *in vivo* using the hydrodynamic tail vein injection assay in mice. Fifty-eight percent (7/12) of high-scoring liver-enhancer predictions yielded robust enhancer activity in the mouse liver, versus zero for the controls (0/5), selected among low-scoring predictions. In summary, promoters often contain unambiguous tissue- and/or cell-specific sequence signatures that can be learned and used for the *de novo* prediction of enhancers.

3.3.3. Cracking the LAnguage of Regulatory Elements (CLARE): a Web-interface for the Discovery of *Cis*-regulatory Elements

To provide and encourage access within the scientific community to our *cis*-regulatory element prediction tools, we set up a web server, which we named CLARE (Cracking the LAnguage of Regulatory Elements, (Taher, Narlikar, et al. 2012)). CLARE is freely available at `http://clare.dcode.org/`.

The only input required from the user is a set of sequences of *cis*-regulatory elements in FASTA format. CLARE proceeds in three main steps:

(i) Randomly sample noncoding regions from the human genome of the same length and and GC-content as the input *cis*-regulatory elements (*control set*). Optionally, the user can upload his/her own control set.

(ii) Search and score putative TF binding sites in the input and control sequences using tf-Search (Ovcharenko, Loots, et al. 2005) with known motifs from the TRANSFAC (Matys, Kel-Margoulis, et al. 2006), JASPAR (Sandelin, Alkema, et al. 2004), and UniPROBE (Robasky and

Bulyk 2011) databases as well as the top 10 overrepresented motifs among the input sequences discovered *de novo* with PRIORITY (Narlikar, Gordân, et al. 2006).

(iii) Build a linear regression classifier.

After running, CLARE returns three primary outputs:

(i) Putative TF binding sites that are relevant to the classification.
(ii) Classifier performance.
(iii) Predictions of *cis*-regulatory elements.

CLARE provides a user-friendly interface for biologists to analyze and prioritize genomic regions and TF binding sites for further experimental validation.

3.4. Parallel Enhancer Testing Suggests a Flexible *Cis*-regulatory Grammar

Despite continual progress in the cataloging of *cis*-regulatory elements, little is known about the grammatical rules that govern their activity. TFs act in a combinatorial and partly redundant manner. Hence, *de novo* creation or deletion of TF binding sites within a particular *cis*-regulatory element can, though not always, modify the intensity and spatiotemporal pattern of expression of its target gene. For example, the gain of new TF binding sites for conserved regulators of wing development within an enhancer that regulates the *yellow* gene resulted in the evolution of a wing spot in *Drosophila biarmipes* (Gompel, Prud'homme, et al. 2005). Deciphering these grammatical rules is essential to enabling high-resolution mapping of *cis*-regulatory elements, accurate interpretation of nucleotide variation within them, and the design of sequences that can deliver molecules for therapeutic purposes in a spatiotemporal manner.

Integrating statistics, combinatorics, and computational techniques we designed a collection of $\sim 5,000$ synthetic *cis*-regulatory element sequences (SRESs) representing binding site arrangements for 12 TFs (AHR/ARNT, CEBPA, FOXA1, GATA4, HNF1A, HNF4A, NR2F2, ONECUT1, PPARA, RXRA, TFAP2C and XBP1) that are known to play an important role in liver development and function (Smith, Taher, et al. 2013). The regulatory activity of the SRESs was examined using a massively parallel reporter assay in the mouse liver (Figure 3.3 A). Briefly, each SRES along with a unique 20-base-pair-long tag was first cloned into an expression vector (pGL4.23) containing a minimal promoter driving transcription of luciferase (minP/luc). The entire sequence of each SRES and its associated tag were determined by sequencing. The library of SRESs was then introduced into three mice through hydrodynamic tail vein injection, livers were harvested after 24 h and sequencing was performed to quantify abundance of transcribed tags. These data were used to estimate the regulatory activity of each SRES. With the aim to test distinct hypotheses regarding the nature of homotypic clustering, synergy between TFs in heterotypic enhancers and the impact of binding site spacing and order on expression, we designed three classes of sequences (Figure 3.3 B):

(i) Class I SRESs ($n = 533$) were homotypic, containing 1, 2, 4 or 8 copies of the same TF binding site with different spacing.
(ii) Class II SRESs ($n = 1,797$) were heterotypic, containing exactly 2 different types of TF binding sites arranged as 2, 4 or 8 sites that were separated uniformly.
(iii) Class III SRESs ($n = 2,636$) were heterotypic, with 3-8 different types of TF binding sites separated by a fixed distance, with only 1 site per TF.

In all cases, the TF binding sites were patterned onto two different 168-base-pair-long sequences which did not exhibit regulatory activity in liver. All of these sequences are enriched in liver-specific enhancers identified using ChIP-seq experiments (Shen, Yue, et al. 2012).

We found that certain TFs act as direct drivers of gene expression in homotypic clusters, independent of spacing, whereas others function only synergistically. Heterotypic enhancers were stronger than their homotypic analogs, and favor specific TF binding site combinations, mimicking putative

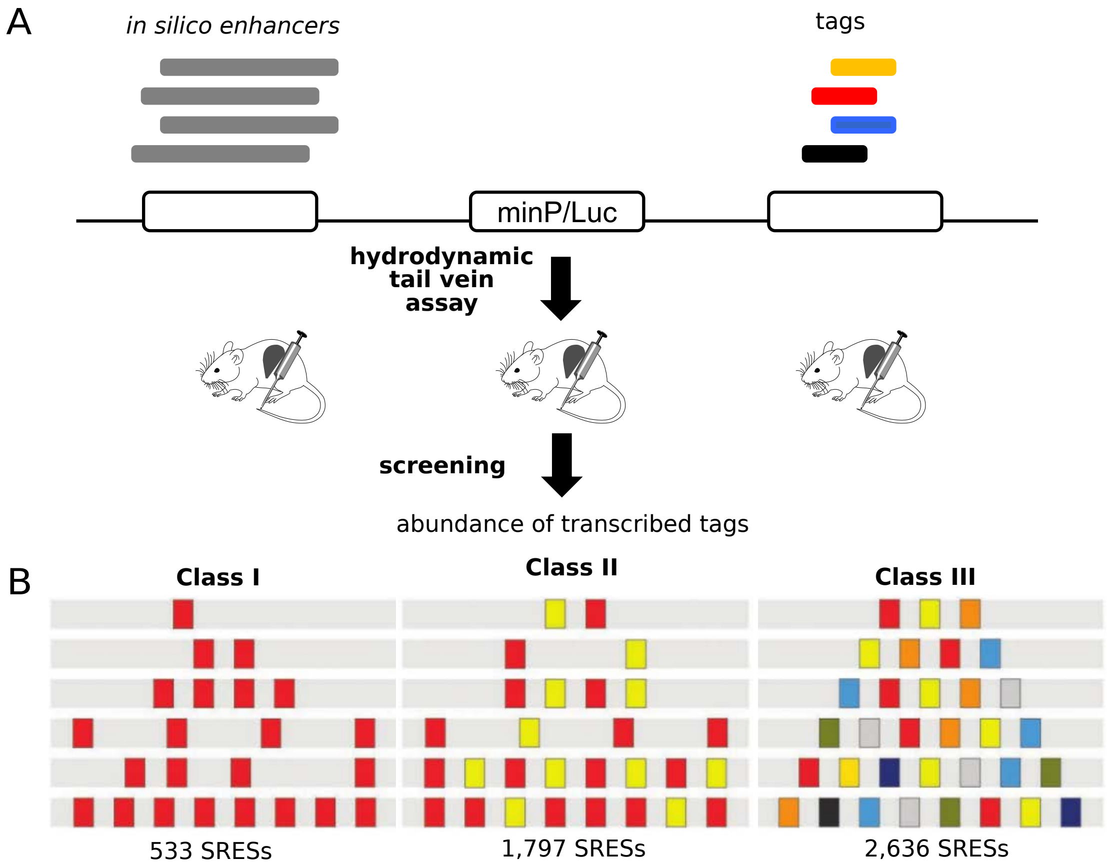

Figure 3.3. Synthetic enhancer sequence design. (A) Schematic of massively parallel reporter assay methodology. SRESs were cloned upstream of a minimal promoter in a tagged luciferase library and then assayed *in vivo* using hydrodynamic tail vein injection. Livers were dissected 24 h after injection, mRNA was generated, and tags were reverse transcribed and sequenced. (B) SRESs consist of patterns of 12 consensus binding sequences arranged homotypically (class I) or heterotypically (class II and class III) on 1 of 2 neutral, 168-bp templates.

native enhancers. Exhaustive testing of TF binding site permutations supported a model with flexibility in binding site order. The flexibility and redundancy of the *cis*-regulatory code can explain both its functional robustness and the apparent simplicity with which changes in *cis*-regulatory elements can alter the spatiotemporal pattern of expression of a gene.

This work provides a unique catalog of tissue-specific synthetic enhancers as well as a massively parallel view of the basic principles of regulatory activity *in vivo*.

3.5. Regulation of Gene Expression in Embryonic Development

During embryonic development, starting with the totipotent zygote, one genome gives rise to multiple cells with different identities. The developmental program requires the coordinated expression of genes encoding TFs and components of cell signaling pathways, and is primarily controlled by TFs binding to specific *cis*-regulatory elements. Developmental regulatory networks are often complex, with multiple levels of cross-talk between different pathways and both positive and negative feedback loops. Detailed information on developmental stage-specific changes in gene expression is crucial for elucidating the regulatory networks underlying development and morphogenesis (Taher, Pfeiffer, et al. 2015).

In an effort to learn more about the genes involved in limb morphogenesis, we employed whole-genome microarrays to examine gene expression across 5 stages of limb development, from limb initiation, at E9.5, to E13.5, when they are fully patterned (Taher, Collette, et al. 2011). Our data and analysis describe the global gene expression dynamics during early murine fore- and hindlimb development, when cartilage, tendons, muscle, joints, vasculature and nerves are specified and the musculoskeletal system of limbs is established. Differential expression was assessed using a linear modeling approach and the empirical Bayes statistics as implemented in the limma R package (Ritchie, Phipson, et al. 2015).

In particular, we observed that the onset of limb formation is characterized by the up-regulation of TFs, which is followed by a massive activation of genes at E10.5 and E11.5. Furthermore, we found that the limb developmental program involves over $3,500$ genes that exhibit ~ 20 distinct expression profiles across the 5 stages considered. Interestingly, approximately 30% of these genes that were identified as significantly up-regulated in the limb were novel, or have not yet been characterized in the limb, dramatically expanding the repertoire of genes that are likely to function in the limb. Hierarchical and stage-specific gene clustering identified expression profiles that are likely to correlate with functional programs during limb development. Further characterization of these transcripts and their regulators will provide new insights into specific tissue patterning processes.

This was, to our knowledge, the first comprehensive analysis of the gene expression dynamics governing limb morphogenesis.

4. Discussion and Outlook

Despite its unprecedented scale, the *cis*-regulatory sequences identified by large international projects such as ENCODE most likely account for only a fraction of all *cis*-regulatory elements in the human genome (He, Kong, et al. 2011). Furthermore, given their cell- and condition-specific nature, the discovery and experimental validation of all regulatory sequences in the human genome is expected to remain an elusive goal for years to come. Recent advances in molecular biology coupled with high-throughput sequencing technologies, including ChIP-seq and DNase-seq, have opened up new possibilities to annotate the noncoding portion of the genome. Nevertheless, the occurrence of biochemical events as reported by these technologies does not necessarily imply *cis*-regulatory activity. In other words, despite their relatively high validation rates, these methods only predict *cis*-regulatory elements. Hence, developing computational methods that can help identify and characterize *cis*-regulatory sequences remains a central problem in biology.

High-throughput sequencing is transforming computational biology. The large datasets generated by current technologies require the development and implementation of appropriate analysis, interpretation, and visualization methods. Unlike coding regions, noncoding regions in the genome typically lacks any annotation regarding their biological functions. We conceived a method to assign biological meaning to a set of noncoding genomic regions by analyzing the annotations of the nearby genes (Taher and Ovcharenko 2009). In addition, to improve our understanding of the language of transcriptional regulation, we established a strategy to ascertain the ancestral identity of diverged noncoding sequences (Taher, McGaughey, et al. 2011). More precisely, we showed that genomic pairwise comparisons among multiple species facilitates the detection of ancestral sequence identity. We applied this principle to determine orthology relationships between thousands of noncoding elements in the human and zebrafish genomes that fail to align under standard pairwise sequence comparison. Moreover, our approach can be easily generalized to other species. For diverged noncoding sequences that defy detection based on sequence similarity, even after including more species into the analysis, we designed an alignment model based solely on the distribution of TF binding sites. Our results demonstrate that *cis*-regulatory elements can maintain their function despite sequence divergence. Further, because of the cell- and condition-specific nature of *cis*-regulatory elements, there is a pressing need for accurate high-throughput approaches that can be used to identify *cis*-regulatory elements in a wide variety of cell types and biological contexts. In contrast to ChIP-seq and DNase-seq, computational approaches provide access to sequence features that contribute to our understanding of transcriptional regulation. We designed different machine learning-based tools that predict *cis*-regulatory elements on the basis of a few characteristics that *cis*-regulatory elements are known to share (Burzynski, Reed, et al. 2012; Busser, Taher, et al. 2012; Taher, Narlikar, et al. 2012; Taher, Smith, et al. 2013; Visel, Taher, et al. 2013). Our results show that distinct combinations of TF binding sites are responsible for cell- and condition-specific regulatory activity. Furthermore, the regulatory code appears to have a relatively simple grammar (Smith, Taher, et al. 2013). Indeed, although there is still no comprehensive picture of the necessary and sufficient sequence features for *cis*-regulatory elements, we have succeeded at creating *cis*-regulatory elements *de novo* from non-functional sequence by the addition of TF binding sites. Interestingly, a single copy of the 17-bp binding motifs for HNF1A or XBP1 is sufficient to drive consistent expression in adult liver when paired with a minimal promoter. To our knowledge, these constitute the shortest functional elements characterized *in vivo* and could, eventually, be used in inducible, as well as non-inducible, transgenic experiments. Finally, we showed how to elucidate transcriptional relationships in the developing embryo from high-throughput gene expression data (Taher, Collette, et al. 2011).

To estimate how our classifiers will generalize on an independent dataset we examined their performance in a cross-validation setting (SVM and linear regression) or their out-of-bag (OOB) error (RF). One round of cross-validation involves partitioning a dataset into two complementary subsets, a *training* and a *test* set. The classifier is subsequently trained on the *training* set and evaluated on the *test* set. Multiple rounds of cross-validation are performed on different partitions to estimate the variability of the results. In the case of RF classifiers, there is no need for cross-validation to get an unbiased estimate of the test set error. Indeed, each tree in a RF is constructed using a different bootstrap sample, consisting in approximately two-thirds of the original dataset. Hence, one-third of the trees can be used as "test" set to evaluate each data point in the dataset. The OOB error is the fraction of the instances in which the class assigned to a given data point is not equal to its true class averaged over all data points. Within these frameworks, we systematically compared our methods to other state-of-the-art approaches. For example, a few approaches were proposed in the past to identify orthologous pairs of *cis*-regulatory elements that do not show any discernible sequence conservation (Berezikov, Guryev, et al. 2004; Blanco, Messeguer, et al. 2006; Hallikas, Palin, et al. 2006). Despite relying on similar models, the optimal parameter configuration of these methods depends on the exact question being addressed. In our framework, they succeeded in retrieving the zebrafish orthologs of the human sequences in less than 20% of the cases, as compared to the 51% achieved by our approach. Also, we compared our machine learning classifiers that predict *cis*-regulatory activity based on sequence features to four state-of-the-art methods: CisModule (Zhou and Wong 2004), Cluster-Buster (Frith, Li, et al. 2003), MSCAN (Alkema, Johansson, et al. 2004), and Stubb (Sinha, Liang, et al. 2006). Our classifiers outperformed all others in terms of both sensitivity and specificity, exhibiting substantially and significantly higher AUCs. Finally, we assessed the quality of our genome-wide predictions according to the following criteria:

- Functional analysis.
- Overlap with DHSs and ChIP-seq data.
- Expression pattern of neighboring genes.
- Experimental validation using reporter gene assays, performed by collaborators.

Reporter gene assays in transgenic zebrafish and mouse provide the most stringent and impartial validation for our computational predictions.

Our models rely on sequence motifs representing TF binding affinities. Several methods have been developed to characterize and predict TF binding affinities. TF binding affinities are then modeled using PWMs, which assume independence between positions, and more sophisticated models. To improve the accuracy of these models, phylogenetic footprinting is often employed to restrict the search to conserved TF binding sites. Despite their simplicity, PWM scores have been shown to be strongly correlated with TF binding affinities (Stormo 2000; Stormo 2013). TF binding affinities are most commonly derived from *in vitro* data, which are context independent. Actual binding is known to be highly dependent on cell-specific conditions, such as chromatin accessibility and TF and coactivator availability (Spitz and Furlong 2012). However, even models integrating multiple types of sequence data achieve only modest accuracies (e.g., (Zhong, He, et al. 2013)). Sequence features alone appear not to be sufficient for functional TF binding site recognition. ChIP-seq measurements of TF binding have been successfully used for *cis*-regulatory element prediction (e.g., (Yip, Cheng, et al. 2012)) and can substitute predicted TF binding sites in a straightforward manner to improve the performance of our tools.

The current state of knowledge suggest that some repetitive sequences in the human genome play an important role in transcription regulation. The regulatory role of repetitive elements such as transposable elements (TEs) was already recognized by Barbara McClintock in the 1940s and 1950s. Nevertheless, TEs have long been dismissed as "junk" DNA, and are only now beginning to receive the attention they deserve. Indeed, $\sim$ 50% of the human genome is derived from repetitive sequences, most of which are classified as TEs (de Koning, Gu, et al. 2011). The proliferation and evolution of TEs have had multiple impacts on the vertebrate genome. For example, TEs have substantially contributed to the expansion of binding sites for CTCF (Schmidt, Schwalie, et al. 2012). Also, DNase I hypersensitivity data from ENCODE has demonstrated that over half of primate-specific open chromatin regions are associated with TEs, and that this association depends on the specific TE family (Jacques, Jeyakani, et al. 2013). However, the repetitive nature of TEs makes them difficult to analyze. Standard analyses of high-throughput sequencing data usually exclude sequences matching to multiple locations of the genome. Including such sequences would improve the sensitivity of the analysis, but at the expense of specificity. Once various technical issues have been addressed, the analysis of extensive panels of epigenetic marks in the near future is likely to show that a large fraction of sequences derived from TEs have regulatory activities.

Although high-throughput whole-genome sequencing has facilitated the identification of noncoding variants, the discrimination of causal mutations for complex diseases is still a formidable challenge. The vast majority of SNPs identified in GWAS reside in noncoding portions of the genome. Indeed, increasing evidence suggests that most causal mutations are expected to lie within *cis*-regulatory elements (Stranger, Stahl, et al. 2011). Moreover, several resources, including HaploReg (`http://www.broadinstitute.org/mammals/haploreg/haploreg.php`), RegulomeDB (`http://regulomedb.org`), and GWAS3D (`http://jjwanglab.org/gwas3d`) have been developed for the specific purpose of annotating noncoding variants from GWAS (Edwards, Beesley, et al. 2013). Nevertheless, medical research continues to focus primarily on protein-coding variants. The main reason for this is our lack of understanding of the general principles of transcriptional regulation and how it controls developmental and disease progression. Hence, an essential future goal is to understand how *cis*-regulatory elements drive specific spatiotemporal patterns of expression and how they interact with each other. The joint analysis of GWAS with functional genomics is key to unleash the value of whole-genome sequencing for personalized medicine.

Bibliography

[1] Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L. A., and Rubin, E. M. "Deletion of ultraconserved elements yields viable mice." *PLoS Biol* 5.9 (2007), e234.

[2] Alkema, W. B. L., Johansson, O., Lagergren, J., and Wasserman, W. W. "MSCAN: identification of functional clusters of transcription factor binding sites." *Nucleic Acids Res* 32.Web Server issue (2004), W195–W198.

[3] Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. "Genome-wide quantitative enhancer activity maps identified by STARR-seq." *Science* 339.6123 (2013), pp. 1074–1077.

[4] Ashburner, M. et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet* 25.1 (2000), pp. 25–29.

[5] Bailey, T. L. and Elkan, C. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." *Proc Int Conf Intell Syst Mol Biol* 2 (1994), pp. 28–36.

[6] Bailey, T. L. and Gribskov, M. "Combining evidence using p-values: application to sequence homology searches." *Bioinformatics* 14.1 (1998), pp. 48–54.

[7] Baylies, M. K. and Bate, M. "twist: a myogenic switch in Drosophila." *Science* 272.5267 (1996), pp. 1481–1484.

[8] Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. "Ultraconserved elements in the human genome." *Science* 304.5675 (2004), pp. 1321–1325.

[9] Bell, A. C., West, A. G., and Felsenfeld, G. "The protein CTCF is required for the enhancer blocking activity of vertebrate insulators." *Cell* 98.3 (1999), pp. 387–396.

[10] Benko, S. et al. "Disruption of a long distance regulatory region upstream of SOX9 in isolated disorders of sex development." *J Med Genet* 48.12 (2011), pp. 825–830.

[11] Benko, S. et al. "Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence." *Nat Genet* 41.3 (2009), pp. 359–364.

[12] Berezikov, E., Guryev, V., Plasterk, R. H. A., and Cuppen, E. "CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting." *Genome Res* 14.1 (2004), pp. 170–178.

[13] Blanchette, M. et al. "Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression." *Genome Res* 16.5 (2006), pp. 656–668.

[14] Blanco, E., Messeguer, X., Smith, T. F., and Guigó, R. "Transcription factor map alignment of promoter regions." *PLoS Comput Biol* 2.5 (2006), e49.

[15] Blow, M. J. et al. "ChIP-Seq identification of weakly conserved heart enhancers." *Nat Genet* 42.9 (2010), pp. 806–810.

[16] Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., and Rubin, E. M. "Phylogenetic shadowing of primate sequences to find functional regions of the human genome." *Science* 299.5611 (2003), pp. 1391–1394.

[17] Breiman, L. *Machine Learning* 45.1 (2001), pp. 5–32.

[18] Bulger, M. and Groudine, M. "Functional and mechanistic diversity of distal transcription enhancers." *Cell* 144.3 (2011), pp. 327–339.

[19] Burzynski, G. M., Reed, X., Taher, L., Stine, Z. E., Matsui, T., Ovcharenko, I., and McCallion, A. S. “Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control.” *Genome Res* 22.11 (2012), pp. 2278–2289.

[20] Busser, B. W., Taher, L., Kim, Y., Tansey, T., Bloom, M. J., Ovcharenko, I., and Michelson, A. M. “A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis.” *PLoS Genet* 8.3 (2012), e1002531.

[21] Buttgereit, D. “Redundant enhancer elements guide beta 1 tubulin gene expression in apodemes during Drosophila embryogenesis.” *J Cell Sci* 105 (Pt 3) (1993), pp. 721–727.

[22] Carroll, S. B. “Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution.” *Cell* 134.1 (2008), pp. 25–36.

[23] Chen, C.-C., Xiao, S., Xie, D., Cao, X., Song, C.-X., Wang, T., He, C., and Zhong, S. “Understanding variation in transcription factor binding by modeling transcription factor genome-epigenome interactions.” *PLoS Comput Biol* 9.12 (2013), e1003367.

[24] Crawford, G. E. et al. “Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites.” *Proc Natl Acad Sci U S A* 101.4 (2004), pp. 992–997.

[25] Cuddapah, S., Jothi, R., Schones, D. E., Roh, T.-Y., Cui, K., and Zhao, K. “Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains.” *Genome Res* 19.1 (2009), pp. 24–32.

[26] de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. “Repetitive elements may comprise over two-thirds of the human genome.” *PLoS Genet* 7.12 (2011), e1002384.

[27] Dickel, D. E. et al. “Function-based identification of mammalian enhancers using site-specific integration.” *Nat Methods* 11.5 (2014), pp. 566–571.

[28] Edwards, S. L., Beesley, J., French, J. D., and Dunning, A. M. “Beyond GWASs: illuminating the dark road from association to function.” *Am J Hum Genet* 93.5 (2013), pp. 779–797.

[29] Enattah, N. S., Sahi, T., Savilahti, E., Terwilliger, J. D., Peltonen, L., and Järvelä, I. “Identification of a variant associated with adult-type hypolactasia.” *Nat Genet* 30.2 (2002), pp. 233–237.

[30] ENCODE Project Consortium. “The ENCODE (ENCyclopedia Of DNA Elements) Project.” *Science* 306.5696 (2004), pp. 636–640.

[31] Epstein, D. J. “Cis-regulatory mutations in human disease.” *Brief Funct Genomic Proteomic* 8.4 (2009), pp. 310–316.

[32] Erwin, G. D., Oksenberg, N., Truty, R. M., Kostka, D., Murphy, K. K., Ahituv, N., Pollard, K. S., and Capra, J. A. “Integrating diverse datasets improves developmental enhancer prediction.” *PLoS Comput Biol* 10.6 (2014), e1003677.

[33] Frith, M. C., Li, M. C., and Weng, Z. “Cluster-Buster: Finding dense clusters of motifs in DNA sequences.” *Nucleic Acids Res* 31.13 (2003), pp. 3666–3668.

[34] Furlong, E. E., Andersen, E. C., Null, B., White, K. P., and Scott, M. P. “Patterns of gene expression during Drosophila mesoderm development.” *Science* 293.5535 (2001), pp. 1629–1633.

[35] Gaszner, M. and Felsenfeld, G. “Insulators: exploiting transcriptional and epigenetic mechanisms.” *Nat Rev Genet* 7.9 (2006), pp. 703–713.

[36] Gompel, N., Prud’homme, B., Wittkopp, P. J., Kassner, V. A., and Carroll, S. B. “Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila.” *Nature* 433.7025 (2005), pp. 481–487.

[37] Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., and Ovcharenko, I. “Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers.” *Genome Res* 20.5 (2010), pp. 565–577.

[38] Gross, D. S. and Garrard, W. T. “Nuclease hypersensitive sites in chromatin.” *Annu Rev Biochem* 57 (1988), pp. 159–197.

[39] Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. “Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity.” *Cell* 124.1 (2006), pp. 47–59.

[40] Hardison, R. C. and Taylor, J. “Genomic approaches towards finding cis-regulatory modules in animals.” *Nat Rev Genet* 13.7 (2012), pp. 469–483.

[41] He, A., Kong, S. W., Ma, Q., and Pu, W. T. “Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart.” *Proc Natl Acad Sci U S A* 108.14 (2011), pp. 5632–5637.

[42] Hébert, J. M. and Fishell, G. “The genetics of early telencephalon patterning: some assembly required.” *Nat Rev Neurosci* 9.9 (2008), pp. 678–685.

[43] Heintzman, N. D. et al. “Histone modifications at human enhancers reflect global cell-type-specific gene expression.” *Nature* 459.7243 (2009), pp. 108–112.

[44] Herz, H.-M., Hu, D., and Shilatifard, A. “Enhancer malfunction in cancer.” *Mol Cell* 53.6 (2014), pp. 859–866.

[45] International Human Genome Sequencing Consortium. “Finishing the euchromatic sequence of the human genome.” *Nature* 431.7011 (2004), pp. 931–945.

[46] Jacques, P.-É., Jeyakani, J., and Bourque, G. “The majority of primate-specific regulatory sequences are derived from transposable elements.” *PLoS Genet* 9.5 (2013), e1003504.

[47] Jankowski, A., Szczurek, E., Jauch, R., Tiuryn, J., and Prabhakar, S. “Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers.” *Genome Res* 23.8 (2013), pp. 1307–1318.

[48] Kelly, B. L. and Locksley, R. M. “Coordinate regulation of the IL-4, IL-13, and IL-5 cytokine cluster in Th2 clones revealed by allelic expression patterns.” *J Immunol* 165.6 (2000), pp. 2982–2986.

[49] King, M. C. and Wilson, A. C. “Evolution at two levels in humans and chimpanzees.” *Science* 188.4184 (1975), pp. 107–116.

[50] Kornberg, R. D. and Lorch, Y. “Chromatin structure and transcription.” *Annu Rev Cell Biol* 8 (1992), pp. 563–587.

[51] Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. “Transposable elements have rewired the core regulatory network of human embryonic stem cells.” *Nat Genet* 42.7 (2010), pp. 631–634.

[52] Kurth, I. et al. “Duplications of noncoding elements 5’ of SOX9 are associated with brachydactyly-anonychia.” *Nat Genet* 41.8 (2009), pp. 862–863.

[53] Lander, E. S. et al. “Initial sequencing and analysis of the human genome.” *Nature* 409.6822 (2001), pp. 860–921.

[54] Landt, S. G. et al. “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.” *Genome Res* 22.9 (2012), pp. 1813–1831.

[55] Lee, T. I. and Young, R. A. “Transcription of eukaryotic protein-coding genes.” *Annu Rev Genet* 34 (2000), pp. 77–137.

[56] Leslie, C., Eskin, E., and Noble, W. S. “The spectrum kernel: a string kernel for SVM protein classification.” *Pac Symp Biocomput* (2002), pp. 564–575.

[57] Levine, M. and Tjian, R. “Transcription regulation and animal diversity.” *Nature* 424.6945 (2003), pp. 147–151.

[58] Lin, Q., Chen, Q., Lin, L., Smith, S., and Zhou, J. “Promoter targeting sequence mediates enhancer interference in the Drosophila embryo.” *Proc Natl Acad Sci U S A* 104.9 (2007), pp. 3237–3242.

[59] Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. “Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.” *Science* 288.5463 (2000), pp. 136–140.

[60] Maniatis, T., Goodbourn, S., and Fischer, J. A. “Regulation of inducible and tissue-specific gene expression.” *Science* 236.4806 (1987), pp. 1237–1245.

[61] Manolio, T. A. et al. “Finding the missing heritability of complex diseases.” *Nature* 461.7265 (2009), pp. 747–753.

[62] Maston, G. A., Evans, S. K., and Green, M. R. “Transcriptional regulatory elements in the human genome.” *Annu Rev Genomics Hum Genet* 7 (2006), pp. 29–59.

[63] Matys, V. et al. “TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.” *Nucleic Acids Res* 34.Database issue (2006), pp. D108–D110.

[64] May, D. et al. “Large-scale discovery of enhancers from human heart tissue.” *Nat Genet* 44.1 (2012), pp. 89–93.

[65] Menke, D. B., Guenther, C., and Kingsley, D. M. “Dual hindlimb control elements in the Tbx4 gene and region-specific control of bone size in vertebrate limbs.” *Development* 135.15 (2008), pp. 2543–2553.

[66] Mikkelsen, T. S. et al. “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.” *Nature* 448.7153 (2007), pp. 553–560.

[67] Mouse ENCODE Consortium et al. “An encyclopedia of mouse DNA elements (Mouse ENCODE).” *Genome Biol* 13.8 (2012), p. 418.

[68] Mouse Genome Sequencing Consortium et al. “Initial sequencing and comparative analysis of the mouse genome.” *Nature* 420.6915 (2002), pp. 520–562.

[69] Narlikar, L. “Multiple novel promoter-architectures revealed by decoding the hidden heterogeneity within the genome.” *Nucleic Acids Res* 42.20 (2014), pp. 12388–12403.

[70] Narlikar, L., Gordân, R., Ohler, U., and Hartemink, A. J. “Informative priors based on transcription factor structural class improve de novo motif discovery.” *Bioinformatics* 22.14 (2006), e384–e392.

[71] Narlikar, L., Sakabe, N. J., Blanski, A. A., Arimura, F. E., Westlund, J. M., Nobrega, M. A., and Ovcharenko, I. “Genome-wide discovery of human heart enhancers.” *Genome Res* 20.3 (2010), pp. 381–392.

[72] Nelson, C. E., Hersh, B. M., and Carroll, S. B. “The regulatory content of intergenic DNA shapes genome architecture.” *Genome Biol* 5.4 (2004), R25.

[73] Ovcharenko, I., Loots, G. G., Giardine, B. M., Hou, M., Ma, J., Hardison, R. C., Stubbs, L., and Miller, W. “Mulan: multiple-sequence local alignment and visualization for studying function and evolution.” *Genome Res* 15.1 (2005), pp. 184–194.

[74] Ovcharenko, I., Stubbs, L., and Loots, G. G. “Interpreting mammalian evolution using Fugu genome comparisons.” *Genomics* 84.5 (2004), pp. 890–895.

[75] Pattabiraman, K. et al. “Transcriptional regulation of enhancers active in protodomains of the developing cerebral cortex.” *Neuron* 82.5 (2014), pp. 989–1003.

[76] Pennacchio, L. A. et al. “In vivo enhancer analysis of human conserved non-coding sequences.” *Nature* 444.7118 (2006), pp. 499–502.

[77] Pennisi, E. “Genomics. ENCODE project writes eulogy for junk DNA.” *Science* 337.6099 (2012), pp. 1159, 1161.

[78] Perry, M. W., Boettiger, A. N., and Levine, M. “Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo.” *Proc Natl Acad Sci U S A* 108.33 (2011), pp. 13570–13575.

[79] Philippakis, A. A., Busser, B. W., Gisselbrecht, S. S., He, F. S., Estrada, B., Michelson, A. M., and Bulyk, M. L. “Expression-guided in silico evaluation of candidate cis regulatory codes for Drosophila muscle founder cells.” *PLoS Comput Biol* 2.5 (2006), e53.

[80] Ren, B. et al. “Genome-wide location and function of DNA binding proteins.” *Science* 290.5500 (2000), pp. 2306–2309.

[81] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. “limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic Acids Res* 43.7 (2015), e47.

[82] Robasky, K. and Bulyk, M. L. “UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions.” *Nucleic Acids Res* 39.Database issue (2011), pp. D124–D128.

[83] Robertson, G. et al. “Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.” *Nat Methods* 4.8 (2007), pp. 651–657.

[84] Roeder, R. G. “The role of general initiation factors in transcription by RNA polymerase II.” *Trends Biochem Sci* 21.9 (1996), pp. 327–335.

[85] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. “JASPAR: an open-access database for eukaryotic transcription factor binding profiles.” *Nucleic Acids Res* 32.Database issue (2004), pp. D91–D94.

[86] Schmidt, D. et al. “Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.” *Science* 328.5981 (2010), pp. 1036–1040.

[87] Schmidt, D. et al. “Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages.” *Cell* 148.1-2 (2012), pp. 335–348.

[88] Schnable, P. S. et al. “The B73 maize genome: complexity, diversity, and dynamics.” *Science* 326.5956 (2009), pp. 1112–1115.

[89] Shen, Y. et al. “A map of the cis-regulatory sequences in the mouse genome.” *Nature* 488.7409 (2012), pp. 116–120.

[90] Siepel, A. et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.” *Genome Res* 15.8 (2005), pp. 1034–1050.

[91] Sinha, S., Liang, Y., and Siggia, E. “Stubb: a program for discovery and analysis of cis-regulatory modules.” *Nucleic Acids Res* 34.Web Server issue (2006), W555–W559.

[92] Smale, S. T. and Kadonaga, J. T. “The RNA polymerase II core promoter.” *Annu Rev Biochem* 72 (2003), pp. 449–479.

[93] Smith, E. and Shilatifard, A. “Enhancer biology and enhanceropathies.” *Nat Struct Mol Biol* 21.3 (2014), pp. 210–219.

[94] Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. “Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model.” *Nat Genet* 45.9 (2013), pp. 1021–1028.

[95] Spitz, F. and Furlong, E. E. M. “Transcription factors: from enhancer binding to developmental control.” *Nat Rev Genet* 13.9 (2012), pp. 613–626.

[96] Staden, R. “Computer methods to locate signals in nucleic acid sequences.” *Nucleic Acids Res* 12.1 Pt 2 (1984), pp. 505–519.

[97] Stormo, G. D. “DNA binding sites: representation and discovery.” *Bioinformatics* 16.1 (2000), pp. 16–23.

[98] Stormo, G. D. “Modeling the specificity of protein-DNA interactions.” *Quant Biol* 1.2 (2013), pp. 115–130.

[99] Strahl, B. D. and Allis, C. D. “The language of covalent histone modifications.” *Nature* 403.6765 (2000), pp. 41–45.

[100] Stranger, B. E., Stahl, E. A., and Raj, T. “Progress and promise of genome-wide association studies for human complex trait genetics.” *Genetics* 187.2 (2011), pp. 367–383.

[101] Stranger, B. E. et al. “Population genomics of human gene expression.” *Nat Genet* 39.10 (2007), pp. 1217–1224.

[102] Su, A. I. et al. “A gene atlas of the mouse and human protein-encoding transcriptomes.” *Proc Natl Acad Sci U S A* 101.16 (2004), pp. 6062–6067.

[103] Su, J., Teichmann, S. A., and Down, T. A. “Assessing computational methods of cis-regulatory module prediction.” *PLoS Comput Biol* 6.12 (2010), e1001020.

[104] Swallow, D. M. “Genetics of lactase persistence and lactose intolerance.” *Annu Rev Genet* 37 (2003), pp. 197–219.

[105] Symmons, O. and Spitz, F. “From remote enhancers to gene regulation: charting the genome’s regulatory landscapes.” *Philos Trans R Soc Lond B Biol Sci* 368.1620 (2013), p. 20120358.

[106] Taher, L., Collette, N. M., Murugesh, D., Maxwell, E., Ovcharenko, I., and Loots, G. G. “Global gene expression analysis of murine limb development.” *PLoS One* 6.12 (2011), e28358.

[107] Taher, L., Narlikar, L., and Ovcharenko, I. “CLARE: Cracking the LAnguage of Regulatory Elements.” *Bioinformatics* 28.4 (2012), pp. 581–583.

[108] Taher, L., Narlikar, L., and Ovcharenko, I. “Identification and computational analysis of gene regulatory elements.” *Cold Spring Harb Protoc* 2015.1 (2015), pdb.top083642.

[109] Taher, L. and Ovcharenko, I. “Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements.” *Bioinformatics* 25.5 (2009), pp. 578–584.

[110] Taher, L., Pfeiffer, M. J., and Fuellen, G. “Bioinformatics approaches to single-blastomere transcriptomics.” *Mol Hum Reprod* 21.2 (2015), pp. 115–125.

[111] Taher, L., Smith, R. P., Kim, M. J., Ahituv, N., and Ovcharenko, I. “Sequence signatures extracted from proximal promoters can be used to predict distal enhancers.” *Genome Biol* 14.10 (2013), R117.

[112] Taher, L. et al. “Genome-wide identification of conserved regulatory function in diverged sequences.” *Genome Res* 21.7 (2011), pp. 1139–1149.

[113] Taylor, M. V. “Muscle development: molecules of myoblast fusion.” *Curr Biol* 10.17 (2000), R646–R648.

[114] Thanos, D. and Maniatis, T. “Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome.” *Cell* 83.7 (1995), pp. 1091–1100.

[115] Thurman, R. E. et al. “The accessible chromatin landscape of the human genome.” *Nature* 489.7414 (2012), pp. 75–82.

[116] Tishkoff, S. A. et al. “Convergent adaptation of human lactase persistence in Africa and Europe.” *Nat Genet* 39.1 (2007), pp. 31–40.

[117] Visel, A., Akiyama, J. A., Shoukry, M., Afzal, V., Rubin, E. M., and Pennacchio, L. A. “Functional autonomy of distant-acting human enhancers.” *Genomics* 93.6 (2009), pp. 509–513.

[118] Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. “VISTA Enhancer Browser–a database of tissue-specific human enhancers.” *Nucleic Acids Res* 35.Database issue (2007), pp. D88–D92.

[119] Visel, A. et al. “A high-resolution enhancer atlas of the developing telencephalon.” *Cell* 152.4 (2013), pp. 895–908.

[120] Visel, A. et al. “ChIP-seq accurately predicts tissue-specific activity of enhancers.” *Nature* 457.7231 (2009), pp. 854–858.

[121] Wagner, T. et al. “Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9.” *Cell* 79.6 (1994), pp. 1111–1120.

[122] Wang, J. et al. “Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.” *Genome Res* 22.9 (2012), pp. 1798–1812.

[123] Wang, Y., Harvey, C. B., Pratt, W. S., Sams, V. R., Sarner, M., Rossi, M., Auricchio, S., and Swallow, D. M. “The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element.” *Hum Mol Genet* 4.4 (1995), pp. 657–662.

[124] Welter, D. et al. “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.” *Nucleic Acids Res* 42.Database issue (2014), pp. D1001–D1006.

[125] Williams, A., Spilianakis, C. G., and Flavell, R. A. “Interchromosomal association and gene regulation in trans.” *Trends Genet* 26.4 (2010), pp. 188–197.

[126] Yip, K. Y. et al. “Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.” *Genome Biol* 13.9 (2012), R48.

[127] Yusufzai, T. M. and Felsenfeld, G. “The 5’-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element.” *Proc Natl Acad Sci U S A* 101.23 (2004), pp. 8620–8624.

[128] Zhong, S., He, X., and Bar-Joseph, Z. “Predicting tissue specific transcription factor binding sites.” *BMC Genomics* 14 (2013), p. 796.

[129] Zhou, Q. and Wong, W. H. “CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.” *Proc Natl Acad Sci U S A* 101.33 (2004), pp. 12114–12119.

[130] Zhu, X. et al. “Differential regulation of mesodermal gene expression by Drosophila cell type-specific Forkhead transcription factors.” *Development* 139.8 (2012), pp. 1457–1466.

A. Eidesstattliche Erklärung

Hiermit erkläre ich eidesstattlich, dass ich die vorliegende Habilitationsschrift selbständig abgefasst und dabei keine fremden, nicht erwähnten Hilfen verwendet habe. Die vorliegende Habilitationsschrift wurde bisher weder im Ausland noch im Inland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde bzw. Fakultät vorgelegt. Ich erkläre, dass ich ein Verfahren zur Erlangung der Habilitation bisher an keiner wissenschaftlichen Einrichtung beantragt habe und mir die Bestimmungen der Habilitationsordnung der Universitätsmedizin Rostock bekannt sind. Außerdem erkläre ich, dass ich bisher kein Habilitationsverfahren erfolglos beendet habe und dass eine Aberkennung einer bereits erworbenen Habilitation nicht vorliegt.

Erlangen, den 30. April 2015

Leila Taher
Department Biologie
Universität Erlangen-Nürnmberg
Staudtstr. 5
91058 Erlangen
`leila.taher@fau.de`