An Evolutionary Approach to Long-Range Regulation

Emmanuel Mongin

Doctor of Philosophy

Department of Human Genetics

McGill University

Montreal

December 2009

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Doctor of Philosophy

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisors Dr. Mathieu Blanchette and Dr. Ken Dewar, who provided me with constant support and guidance from the beginning until the completion of this thesis. I am deeply grateful to Mathieu Blanchette for integrating me into his working group; the daily interaction with so many brilliant people clearly helped me solve many problems. A great thank to my supervisors for reviewing my manuscript. I also wish to thank Dr. Rob Sladek as member of my advisory committee who largely contributed to this thesis with its insightful advices and comments.

Part of my PhD work was carried out at the EMBL Heidelberg. Special thanks go to Dr. Jochen Wittbrodt who accepted me in his laboratory and to Dr. Laurence Ettwiller, with whom I share the credit on the project of the characterization of *cis*-regulatory regions. I am also grateful to Jochen Wittbrodt's lab members, Thomas Auer and Franziska Gruhl, for their great help.

I also wish to thank Dr. Francois Spitz who gave me many insightful comments and ideas and Pablo Cingolani for his work on the *cis*-regulatory modules website. Sincere appreciation goes to my collegues and friends Dr. Abdoulaye Banire Diallo, Vincenzo Forgetta for their support and Dr. David Serre for giving me its advice on part of the manuscript. I am grateful to Genome Quebec and Genome Canada for funding my PhD project as well as to the EMBL Short-Term Travel Fellowship for funding my trips to Germany.

I also wish to express all my gratitude to my friends, especially Dr. Jer-Ming Chia, Dr. Bruno Lansard and Geneviève La whose support and encouragements were priceless.

I am enormously indebted to my parents, Dr. Christian Mongin and Isabelle Mongin, for their unconditional support through all of these years and the great educational opportunities they provided me with. Last but not least, I express all my thankfulness to my wife, Shuni Tsou, for her patience, assistance and love.

ABSTRACT

Long-range regulatory regions play important functions in the regulation of transcription and are particularly involved in the precise spatio-temporal expression of target genes. Such regions have specific characteristics, among which is their ability to regulate many target genes that can be located up to 1Mb from the transcription start site. The prediction and functional characterization of such regions remains an open problem. Evolutionary approaches have been developed to detect regulatory regions that are under purifying selection. However, little has been done with regards to the impact of long-range regulation on genome evolution.

This thesis focuses on three different aspects of long-range regulation: i/ First we develop a method that predicts regions particularly prone to the fixation of evolutionary breakpoints. We discuss the results obtained in the context of long-range regulation and show that this type of regulation is a major factor shaping vertebrate genomes in evolution. ii/ The second project aims at predicting functional interactions between regulatory regions and target genes based on the observation of evolutionary rearrangements in various vertebrate species. We show how this approach produces a biologically meaningful prediction dataset that will be useful to researchers working on regulation. iii/ Third, we focus on the *in vivo* characterization of regulatory regions. We present a powerful and reliable enhancer detection pipeline composed of an *in silico* approach to predict putative enhancers

and an *in vivo* method to functionally characterize the expression specificity of predicted regions in the developing medaka fish.

The results presented in this thesis contribute to different areas of research such as a better understanding of evolutionary dynamics related to evolutionary rearrangements and to a better *in silico* and *in vivo* characterization of *cis*-regulatory regions.

ABRÉGÉ

La régulation longue distance a d'importantes fonctions dans la régulation de la transcription et est particulièrement impliquée dans la régulation spatiale et temporelle des gènes cibles. Ces régions ont des caractéristiques spécifiques telles que la capacité de contrôler different gènes à des distances jusqu'a 1Mb du site d'initiation de la transcription. La prédiction et la caractérisation fonctionelle de ces regions restent un problème d'actualité. Des approches évolutionaires ont été développées pour détecter les régions sous pression de sélection. En revanche, peu a été fait en rapport avec l'impact de la régulation de longue distance sur l'évolution du génome.

Cette thèse se concentre sur trois différents aspects de la régulation longue distance: i/ Premièrement, nous developpons une méthode de prédiction des régions particulièrement sujettes à la fixation des réarrangements de l'évolution. Nous étudions les résultats obtenus dans le contexte de la régulation longue distance et nous montrons que ce type de régulation est un composant majeur dans le façonnement du génome au cours de l'évolution. ii/ Le second projet à pour but de prédire les interactions fonctionnelles entre les régions de régulation et leur gènes cible à partir de l'observation de réarrangements de l'évolution dans différentes espèces. Nous montrons comment une telle approche produit des résultats biologiquement significatifs qui seront particulièrement utiles aux chercheurs travaillant dans le domaine de la régulation. iii/ Troisièmement, nous nous concentrons sur la caractérisation fonctionnelle *in vivo* des régions

régulatrices. Nous présentons une méthode fiable de détection des enhancers composée d'une approche informatique pour la prédiction de ces régions et d'une approche biologique pour caractériser fonctionnellement les spécificités d'expression de ces régions dans le poisson medaka.

Les résultats présentés dans cette thèse contribuent à une meilleure compréhension des dynamiques d'évolution en relation avec la régulation longue distance et une meilleure prédiction et caractérisation fonctionnelle de ces régions régulatrices.

PREFACE

Contribution to original knowledge

This thesis is written in the form of a manuscript based thesis. It contains three papers that correspond to **Chapters 2, 3, and 4**.

- Chapter 2:

  Emmanuel Mongin, Ken Dewar, Mathieu Blanchette. Long-range regulation is a major driving force in maintaining genome integrity. *BMC Evol Biol*, 9(1):203, Aug 2009.

- Chapter 3:

  Emmanuel Mongin, Ken Dewar, Mathieu Blanchette. Mapping association between long-range *cis*-regulatory regions and their target genes using synteny. *Manuscript in preparation.* To be submitted to PLoS Computational Biology in January 2010.

- Chapter 4:

  Emmanuel Mongin*, Thomas Auer*, Frank Bourrat, Franziska Gruhl, Ken Dewar, Mathieu Blanchette, Jochen Wittbrodt, Laurence Ettwiller. A new molecular tool for dissecting the developing vertebrate nervous system. *Submitted to Nature Methods.* * co-first authors.

Contribution of authors

All of the work presented in Chapter 2 and 3 has been done by me under the supervision of Mathieu Blanchette and Ken Dewar. Mathieu Blanchette revised the manuscripts. The work presented in Chapter 4 is a collaborative project

with Jochen Wittbrodt's laboratory at the EMBL (Heidelberg, Germany). I and Thomas Auer share the co-authorship on this paper. The contributions of the authors are as follow: EM and LE designed the experiments. EM undertook the module prediction as well as the bioinformatic analysis of the data (except location analysis, LE). LE and TA prepared the constructs. EM, LE, TA, FG injected the embryos. TA tested the endogenous minimal promoters and did the *in situ* experiment. FB annotated the embryos. LE prepared the transgenic lines, took and prepared the pictures presented in this thesis. EM and LE shared most of the writing of the manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Gene regulation in higher eukaryotes

The regulation of transcription (and more specifically initiation of transcription), is the main mechanism responsible for the accurate development of eukaryote organisms, tissue specificity, proper response to external stimuli or cell cycle control. In this section, we describe the basic mechanisms of transcriptional regulation as well as the different post-transcriptional mechanisms that also influence the final product of the gene but that will not be discussed further in the thesis.

1.1.1 Overview of transcriptional regulation

Eukaryotic transcription takes place in the nucleus and is generally defined as the synthesis of ribonucleic acid (RNA), or more specifically messenger RNA (mRNA), from a deoxyribonucleic acid (DNA) template. This process requires three main steps: i/initiation of transcription, ii/ elongation, and iii/ termination. Transcription initiation - the main step that controls the expression of the gene - is triggered by proteins named transcription factors (TFs), which bind specific regions of the DNA or *cis*-regulatory regions. Transcription factors can be functional by themselves or may need co-activators in order to be functional. They usually contain a DNA binding domain that is responsible for binding the DNA as well as one or more trans-activating domains with binding affinities to

other transcription factors [Ptashne, 1988]. In some cases, to reach a functional state, transcription factors need to be activated. A TF can be activated through different processes such as: i/ ligand-activated transcription factors (e.g. nuclear hormone receptors [Fondell *et al.*, 1996]), phosphorylation (e.g STATs transcription factors [Lodish *et al.*, 1995], p 916) or interaction with other TFs (e.g. p300 [Eckner *et al.*, 1994]).

The synthesis of RNA from DNA is mediated by a protein complex called RNA polymerase. Its enzymatic activity was first reported by Weiss and Gladstone [Weiss & Gladstone, 1959]. Since then, three different types of RNA polymerases have been described: i/ RNA polymerase I is responsible for the transcription of most ribosomal RNAs (rRNAs), ii/ RNA polymerase II is responsible for the transcription of all protein-coding genes as well as some non-coding RNAs such as microRNAs (miRNA), iii/ RNA polymerase III synthesizes transfer RNA (tRNAs), 5S rRNAs and various other non-coding RNAs (ncRNA). For more general information, see [Lodish *et al.*, 1995], p 405-480.

1.1.2 Overview of post-transcriptional regulation

After transcription, the mRNA (termed pre-mRNA at this stage) undergoes several processing steps that produce a mature mRNA. Those steps combined with regulation of transcription and epigenetic effects, will affect the final concentration of protein or RNA in the cell.

Polyadenylation. All mRNAs (except histone mRNAs), have 3' poly-A tail. The poly-A tail is involved in different post-transcriptional regulatory mechanisms such as mRNA stability, mRNA translational efficiency and transport. The poly-A

tail is added to the pre-mRNA by a combination of i/ endonucleotic cleavage at a specific site (the polyadenylation signal sequence), followed by ii/ the poly A synthesis [Colgan & Manley, 1997]. The poly-A tail directly impacts the stability of the mRNA by providing binding site to the poly(A) binding protein that plays a role in mRNA stability [Coller *et al.*, 1998] (the absence of poly-A tail results in its degradation). The poly A tail also impacts mRNA's translational efficiency and its transport (reviewed in [Wickens *et al.*, 1997, Colgan & Manley, 1997]). Polyadenylation is linked to the presence of polyadenylation signal sequences, and the absence or presence of alternate sites can directly affect the stability of the mRNA or its translation [Beaudoing & Gautheret, 2001].

microRNAs. Another level of post-transcriptional regulation is conducted by short (21-23 nucleotides) single strand molecules, microRNAs (miRNAs). miRNAs, which were first described in the nematode *Caenorhabditis elegans* [Lee *et al.*, 1993, Wightman *et al.*, 1993], bind the 3' untranslated region (UTR) of specific mRNAs by complementarity and promote faster degradation of the mRNA or inhibit translation (reviewed in [Bushati & Cohen, 2007]). It is estimated that about one third of eukaryotic genes may be regulated by miRNAs [Lewis *et al.*, 2005].

Alternative splicing. Alternative splicing (AS) is a biological process that ligates exons of a pre-mRNA in different manners. Constitutive alternative splicing is part of the normal pre-mRNA processing and results in the production of different isoforms and greatly augments the number of different proteins and RNAs that are potentially encoded by a single gene. AS affects at least 74% of

human genes with more than one exon [Johnson *et al.*, 2003]. However, AS can also affect the level of the final product by either integrating stop codons in the coding sequence [Saltzman *et al.*, 2008] or producing isoforms with different targets for miRNAs [Tan *et al.*, 2007]. Alternative splicing is also tissue-specific, therefore certain isoforms are specifically formed in certain cell types and not others.

RNA editing. RNA editing is a post-transcriptional process that involves the deanimation of nucleic acids (cytosines to uracils and adenosines to inosines). RNA editing is directly involved in post-transcriptional regulation by potentially affecting splice site sequences or RNA degradation [Agranat *et al.*, 2008, Weber *et al.*, 2007].

Regulation of translation. Translation can be separated into four steps; initiation, elongation, termination and ribosome recycling [Sonenberg & Hinnebusch, 2009]. The most important step concerning regulation of translation is initiation. Translation initiation starts with the binding of the pre-initiation complex (PIC), which binds the 5' UTR end of the mRNA and starts tracking for the AUG start codon. This binding is activated by eukaryotic initiation factors (eIF) that recognize either the 5' or the 3' end of the mRNA. The initiation of translation is dependent on the presence of these eIF and in case of stress or starvation, the cell inhibits these proteins. miRNA may also recruit eIF inhibitor and so affect translation initiation (translation initiation mechanisms are reviewed in [Sonenberg & Hinnebusch, 2009]). For more general information on post-transcriptional regulation, see [Lodish *et al.*, 1995], p 485-540.

1.2 Mechanisms involved in the regulation of transcription

1.2.1 Chromatin compaction

In eukaryotes, the initiation of transcription depends on the precise binding of trans-acting elements (usually proteins) to a set of specific *cis*-regulatory regions. Elements such as activators, chromatin modifying enzymes and general transcription factors are recruited into an ordered manner to correctly regulate a target gene depending on the cell environment [Cosma, 2002].

The most primary way to control the expression of the gene is to block (or permit) the access of transcription factors to the DNA. DNA in the cell is not linear but is associated with proteins and organized into a higher structure called chromatin. The nucleosome, the fundamental unit of the chromatin, is composed of 147 base pairs (bp) of DNA folded around a histone protein and is directly involved in the compaction of chromatin. Chromatin is organized into different levels of compactions that directly influence the propensity of the gene to be transcribed. Compaction of chromatin can be regulated by histone modifications (methylation and acetylation), DNA methylation and DNA binding protein (e.g. transcription factors). Low level of compaction allows transcription factors to bind directly the DNA, whereas high compaction level would prevent those transcription factors from accessing the DNA and triggering initiation of transcription. Therefore, these epigenetic processes involved in chromatin compaction are the first level of gene regulation.

Histone modification. Histones are targets for covalent modifications such as methylation and acetylation, modifications that change their interaction with

DNA. Those modifications that are linked to gene activation or silencing are described as the "histone code" [Cosma, 2002]. For example, histone H3 K9/K14 diacetylation and H3 K4 trimethylation are associated with transcribed genes, whereas H3 K27 trimethylation are linked to gene silencing [Roh *et al.*, 2006]. Moreover, nucleosomes have higher affinity to certain DNA sequences and are consequently not distributed evenly on the genome, which can affect transcription initiation [Segal *et al.*, 2006].

DNA methylation. DNA methylation is another element of the "epigenetic code" that generates a silent chromatin state [Jaenisch & Bird, 2003]. Cytosine methylation is limited to CpG dinucleotides in mammals. Studies of such modifications are an important source of information to understand which region of chromatin is in an active state. CpG dinucleotides are depleted from the genome and are usually found in the form of CpG islands. CpG islands are regions over 200 bp (upstream of the TSS), with over 50% GC content, and with an observed/expected CpG ratio over 60% [Gardiner-Garden & Frommer, 1987]. Those short genomic regions are enriched in promoter sequences and overlap approximately with 50% of mammalian promoters [Antequera & Bird, 1993, Ioshikhes & Zhang, 2000]. The level of repression of a gene seems to be related to the density of methyl cytosines in the promoter, which is directly linked to the number of CpG regions in the promoters. Therefore, methylation of high density CpG promoters can repress strong promoters whereas low densities of CpG island only repress weak promoters [Bird, 1995]. More recent studies have uncovered that CpG rich promoters

are actually hypo-methylated and is probably not the main repression process [Rollins *et al.*, 2006, Weber *et al.*, 2007].

In eukaryotic organisms, transcription factors target specific regions of the genome (given that the chromatin is in an open state) which have distinct functions in the regulation of transcription initiation. These regions are divided into three main classes: i/ the core promoter, ii/ the proximal promoter , and iii/ long-range regulatory regions. We provide an overview of these different regions, although most of the work involved in the context of this thesis is based on long-range regulatory regions.

1.2.2 Core promoter

The core promoter is directly involved in the fundamental stage of transcription initiation - the assembly of the pre-initiation complex (PIC) which is responsible for the initiation of transcription. The core promoter is located +/- 35 bp from the first transcribed nucleotide or transcription start site (TSS) [Smale & Kadonaga, 2003], but as will be discussed later, this view has been redefined by recent studies. Different types of core promoters exist and are mainly defined by the type of elements they contain and the type of genes they regulate. Among those regions, we find: i/ TATA box, located from 28 to 34 bp upstream from the TSS. TATA box, contrary to previous beliefs, are only found in a small subset of promoters. A recent study showed that only 16% of promoters with measured expression had a TATA box [Cooper *et al.*, 2006]. ii/ Initiator element (Inr) is usually associated with the TATA box and both are sufficient to recruit

the PIC and initiate transcription [Sandelin *et al.*, 2007]. iii/ Downstream promoter element (DPE), located 28-32 bp downstream from the TSS and usually found in TATA-less promoter, seems to have a function similar to TATA by being involved in PIC recruitment [Kadonaga, 2002]. iv/ TFIIB recognition elements, found in both TATA and TATA-less promoters, are involved in transcription modulation. Finally, although, not physically included within the core promoter, the presence of CpG islands in the vicinity of the TSS directly depends on the type of promoter. Core promoters that are close to CpG islands are usually TATA-less promoters and usually associated to genes with housekeeping functions [Carninci *et al.*, 2006, Sandelin *et al.*, 2007]. Core promoters are under the influence of long-range enhancers and there is evidence that certain enhancers would interact specifically with certain promoters. Long-range enhancers are binding regions for transcription factors that are involved in gene transcriptional regulation and that can be located as far as 1Mb from the gene they regulate (see Section 1.2.4 for more details). Butler *and* Kadonaga showed that activation specificity of the enhancers they tested was dependent on the presence of TATA box or DPE motifs. Some enhancers would only interact with TATA promoter whereas others only with DPE motif containing promoters [Butler & Kadonaga, 2001].

Typically, PIC assembly (which contains among various proteins the RNA polII) is initiated by TFIID that binds specific sequences such as the TATA box or the initiator element sequence. General transcription factors (GTFs) subsequently bind the complex and/or the DNA and transcription is initiated (general transcription machinery is reviewed in [Thomas & Chiang, 2006]). Additional regulatory

elements are needed in order to confer an enhanced and more specific expression pattern to the gene. Mediators bind to the general transcription complex with the purpose of conveying more specific regulatory signal to the transcription machinery by interacting with activators [Hampsey & Reinberg, 1999, Cosma, 2002]. Co-activators do not have DNA binding properties but bind to activator protein and link to the PIC [Maston *et al.*, 2006]. PIC assembly can occur only if the chromatin is in an open state configuration and should be put in the context of chromatin modification described previously.

Large scale analysis of transcription start sites studies benefited from a new technology called cap analysis gene expression (CAGE). CAGE is based on the isolation of the 20 first nucleotides from the 5' end of full length cDNAs and their sequencing [Shiraki *et al.*, 2003]. These sequences mapped to the genome allow the localization of transcription start site. CAGE methodology participated in redefining the localization of the transcription start site [Carninci *et al.*, 2005, Carninci *et al.*, 2006]. In this new model, the promoter does not contain one fixed TSS, but usually a collection of TSS located over a 50-100bp window [Sandelin *et al.*, 2007]. However, the distribution of TSS apparently depends on the type of promoters, and the TATA promoter has usually a more precise TSS; whereas promoters associated with CpG islands show a wider distribution of their TSS [Sandelin *et al.*, 2007].

1.2.3 Proximal promoter

The distinction between proximal promoter and core promoter is small and many studies do not make such a distinction. Here we define the proximal promoter as the regions located within 1.5 kb from the TSS (except for the basal promoter already described), which contain binding sites for transcription factors that confer specificity to the expression of the gene. Transcription factors binding these regions are close enough to the TSS to interact directly with the transcription initiation complex without facilitating mechanisms [Bondarenko *et al.*, 2003]. Certain types of activators that bind the proximal promoter trigger the recruitment of the transcription initiation complex [Ptashne & Gann, 1997]. However, this model works only if activator binding regions and basal promoters are located in the same vicinity [Bondarenko *et al.*, 2003], therefore at distances over 1.5 kb, specific facilitating mechanisms are necessary for facilitating interaction between regulatory proteins which bind long-range *cis*-regulatory regions and target genes. A *cis*-regulatory region located over 1.5 kb from the TSS is considered in the context of this thesis as a long range regulatory region. Proximal promoters usually are involved in the regulation of one gene although some of them, named bidirectional promoters can initiate the transcription of gene pairs arranged in opposite strands [Trinklein *et al.*, 2004] (See [Yang & Elnitski, 2008] for an example of prediction method of bidirectional promoters).

1.2.4 Long-range regulatory region

As discussed above, specific mechanisms are needed to facilitate the interaction between transcription factors binding far from the TSS and the core promoter.

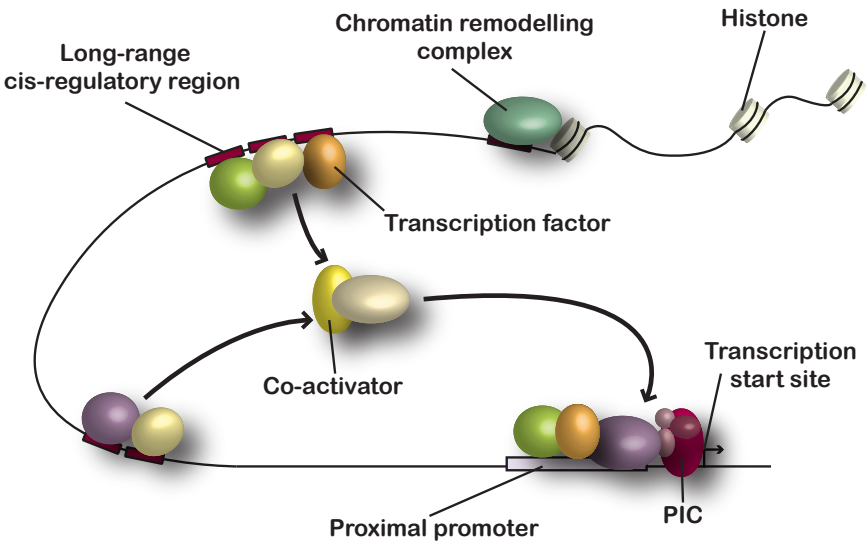

Figure 1–1: ***Cis*-regulatory region in eukaryotic genomes.** Figure adapted from [Wray *et al.*, 2003] and [Wasserman & Sandelin, 2004]. Long-range regulatory regions, proximal promoter and core promoter in the context of genomic DNA.

Among the proposed mechanisms, a looping model where bending properties of the DNA allow the activators to interact with the core promoter is probably facilitated by the DNA supercoiling properties, which favor interaction between long-range enhancers and proximal promoter (on distances over 2,5 kb) [Liu *et al.*, 2001]. See Figure 1–1 for long-range *cis*-regulatory regions in the context of looping DNA. In another model called the tracking mechanism, the *cis*-regulatory region acts as a platform where transcription factors bind and start tracking the core promoter (those different mechanisms are reviewed in [Bondarenko *et al.*, 2003]).

Long-range regulatory regions have distinct properties. First, they regulate the transcription of genes over large distances. For example, the Shh gene is regulated from a region located as far as 1 Mb from its TSS [Sagai *et al.*, 2005]. Contrasting with proximal promoter, a single regulatory region can affect the expression of more than one gene. However, the distance of the gene to the *cis*-regulatory region is important, since genes that are closer to activating *cis*-regulatory regions are more efficiently competing for activation [Dillon *et al.*, 1997]. Finally, long-range regulatory regions usually have the ability to function in an orientation-independent manner.

Within regulatory regions, transcription factor binding sites tend to cluster together [Arnone & Davidson, 1997, Howard & Davidson, 2004] and are organized in regions that have been termed *cis*-regulatory modules (CRMs). It was indeed observed that many putative regulatory regions show an over-representation of binding sites for a limited number of TFs over a restricted region of the genome. Binding sites are arranged in homo-typic or hetero-typic clusters and usually work

cooperatively to trigger transcription [Ptashne, 1988, Hannenhalli & Levy, 2002, Yu *et al.*, 2006].

Since long-range regulators can regulate genes over large distances, mechanisms are required to block the action of enhancers in regions where such regulation is not needed, creating independent unit of regulation [Kellum & Schedl, 1991]. This is performed by regions called insulators which come in two types: i/ Insulators that block the action of an enhancer over a certain region by interacting with another insulator and creating a DNA loop [Farrell *et al.*, 2002] or by blocking the tracking enhancer on linear DNA [Geyer & Corces, 1992, Bondarenko *et al.*, 2003]. ii/ Insulators that prevent the spread of heterochromatin preventing gene silencing [Sun & Elgin, 1999].

Enhancers and silencers have commonly been referred to as long-range regulators. They share many characteristics such as position, orientation independence [Atchison, 1988], and tissue specificity. There is evidences that enhancers under certain conditions can act as silencers [Burke & Baniahmad, 2000]. In the context of this thesis, we will use the general term long-range *cis*-regulatory regions to describe *cis*-elements, including enhancers and silencers, involved in regulation. However, in certain contexts such as *in vivo* characterization of *cis*-regulatory region, we will specifically refer to enhancer elements.

1.3 Experimental *cis*-regulatory region detection

We describe in this section various *in vitro* methods to characterize regions involved in regulation of transcription. We will discuss more specifically about *in vivo* methods in a further section.

1.3.1 Transcription factor binding site identification

The DNA binding domains of transcription factors interact with specific DNA sequences with specific biochemical properties. Different combinations of DNA strings may have similar biochemical properties, therefore transcription factors can bind diverse sequences. In addition to being degenerated, transcription binding sites are short; from 5 to 15 bp. Finally, a transcription factor can adopt different conformations depending on the binding of ligands or co-factors to the transcription factor. The identification of transcription factor binding sites either *in vivo* or *in vitro* is a difficult task and several *in vitro* methods have been developed to capture DNA sequences involved in binding transcription factors.

Systematic Evolution of Ligands by Exponential Enrichment (SELEX) was developed in 1990 [Ellington & Szostak, 1990, Pollock & Treisman, 1990]. The aim of this method is to characterize ligand binding specificity from large pool of random oligonucleotides (DNA or RNA). These oligonucleotides are exposed to ligands such as transcription factors. Bound sequences are subsequently retrieved and characterized by sequencing (reviewed in [Klug & Famulok, 1994]). The main advantage of this method is that no prior knowledge is required about the binding specificity of the transcription factor. The pool of DNA synthesized (about 10^{15}) is large enough to produce sequence with binding specificity to the transcription factor, however, since the experiment is undertaken *in vitro*, it does not reflect *in vivo* conditions.

Electrophoretic Mobility Shift Assay (EMSA) is another method whose principle is based on the property that a complex, formed by a DNA sequence and

a protein, runs slower on a gel electrophoresis than the protein or the DNA fragment alone [Fried & Crothers, 1981]. A double-stranded DNA sequence known to contain the binding site under study is labelled (by fluorescence or radioactivity) and exposed to nuclear extract or purified proteins [Chorley *et al.*, 2008]. If the DNA sequence binds the protein, the complex will run slower on a gel and the binding is therefore detected. This method is particularly sensitive and therefore useful to detect difference of binding affinity between binding sites with different mutations (such as point mutations). However, this method cannot be applied for large scale identification of binding site and as mentioned for SELEX, *in vitro* conditions do not always reflect biological conditions.

1.3.2 Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) is a method to detect binding specificity of transcription factors over genomic DNA. The method is based on the fixation of transcription factors on DNA by formaldehyde crosslinking and shearing. Fixed transcription factors are subsequently precipitated and recovered with antibodies specific to the transcription factor(s) under study [Orlando, 2000]. DNA fragments attached to the precipitated TFs are subsequently identified by sequencing. ChIP technology provides evidence for an association of a transcription factor with a certain region of the chromatin *in vivo* at a certain time and in a certain tissue. The cloning and sequencing steps are particularly time consuming, which prevents the use of basic ChIP technology for large scale studies. This limitation was circumvented by combining ChIP method with DNA microarray (now termed ChIP-on-CHIP). This was first applied to study the localization of Gal4 and Ste12

transcription factors in the yeast genome. Bound DNA sequences were characterized with a microarray containing yeast intergenic sequences [Ren *et al.*, 2000]. This technology was subsequently applied to yeast genomes for specific transcription factors [Ren *et al.*, 2000, Lieb *et al.*, 2001, Iyer *et al.*, 2001] and later for most yeast transcription factors [Lee *et al.*, 2002, Harbison *et al.*, 2004]. With the development of high-density-tiling oligo arrays, studies on whole human chromosomes became possible. For example, NF-κB [Martone *et al.*, 2003] was mapped on chromosome 22, CREB on chromosome 22 [Euskirchen *et al.*, 2004] and Sp1, cMyc, and p53 on chromosome 21 and 22 [Cawley *et al.*, 2004]. But these methods are still limited by the density of oligo arrays. Alternatively, Wei *et al.* [Wei *et al.*, 2006] developed a method were the ChIP step is combined with paired-end ditag (PET) sequencing (ChIP-PET), technology that allowed them to map p53 transcription factor to the whole human genome (ChIP-PET reviewed in [Fullwood *et al.*, 2009]).

1.3.3 DNA methylation profiling

DNA methylation plays an important role in transcriptional regulation. Therefore, the identification of the DNA methylation state in different cell types and conditions is an important component in understanding epigenetic impact on transcriptional regulation. The most established method is the bisulfite DNA sequencing method [Frommer *et al.*, 1992]. This method is based on a sequencing technology that uses bisulfite induced modifications on genomic DNA, so cytosine is turned into uracil whereas 5-methylcytosine is unmodified [Frommer *et al.*, 1992]. Combined with polymerase chain reaction (PCR) and

sequencing, methylated regions can be characterized. Such approach was used on large genome sequences to characterize the methylation profiling on 3 different human chromosomes within various cell types [Eckhardt *et al.*, 2006].

1.3.4 DNAse I hypersensitive sites

The accessibility of *cis*-acting DNA sequence to transcription factors depends on the degree of compaction of the chromatin [Gross & Garrard, 1988]. Wu *et al.* [Wu *et al.*, 1979] first recognized that nucleosome free chromatin is particularly sensitive to nucleases. They subsequently developed a method based on the digestion of chromatin by DNAse I followed by labeling with a probe of the resulting DNA fragments [Wu, 1980]. This method has been applied in the last decades but it somewhat time consuming and inaccurate. More recently, a study associated DNase digestion with massive parallel signature sequencing (MPSS) [Crawford *et al.*, 2006]. Applied on $CD4^+$ T cells, they identified about 14,200 DNase hypersensitive sites. This new approach, which can be applied at the genome-scale, opens new doors to study chromatin state in various cell lines and conditions.

1.3.5 Chromosome Conformation Capture

Although not strictly a method to detect *cis*-regulatory regions, Chomosome Conformation Capture (3C), developed by Dekker *et al.* [Dekker *et al.*, 2002, Dekker, 2003], detects regions of chromatin that interact together and is of relevance to this thesis. This method, based on formaldehyde induced cross-link of regions in contact, is followed (after digestion and ligation) by a PCR amplification of regions that were in contact. This method is also able to retrieve the frequency

with which two sites are in contact. So this approach is particularly adapted to verify experimentally if two DNA-bound proteins specific to two regions such as a promoter and a long-range enhancer are in physical contact. Following the publication of this procedure, improvements were brought to the method, especially to allow large scale analysis of such interactions. With technics such as Chromosome Conformation Capture-on-Chip (4C) [Simonis *et al.*, 2006] and Chromosome Conformation Capture Carbon Copy (5C) [Dostie *et al.*, 2006, Fraser *et al.*, 2009], protocols have been developed to characterize PCR products on microarrays and/or by high throughput sequencing. Even though the observation of an interaction between two loci does not mean that there is a function [Dekker, 2008], such interaction map is particularly useful to confirm predictions of functional association between *cis*-regulatory modules and genes.

1.4 Computational *cis*-regulatory region detection

In the previous section, we described experimental methods to detect and characterize regulatory regions. In this section, we detail various *in silico* approaches to detect genomic regions that control transcription processes. Identifying such regions is a particularly challenging task. Contrasting with protein-coding region predictions, no canonical sequence signatures, such as codon usage or splice site patterns for coding regions, can be used to identify *cis*-regulatory regions. Moreover, *cis*-regulatory regions are not specifically located in particular regions of the genome. Consequently, different strategies were developed and often combined with each other to detect those regions with regulatory potential.

1.4.1 Binding site prediction

With experimental methods such as EMSA, SELEX or ChIP, DNA binding sites for specific transcription factors can be determined. From these sequences, it is possible to calculate the frequency for each DNA nucleotide at each position of the binding site. This information can be represented by position weight matrices (PWMs), also known as position specific score matrices (PSSM) [Stormo, 2000].

Given a set of sequences N bound by a specific transcription factor, the probability $p(b, i)$ to observe a base b at a position i of the sequence is shown equation 1.1 where $f_{b,i}$ is the count for base b at position i, $s(b)$ and $s(b')$ are pseudocount functions. A pseudocount function is used to add a negligible value to observed data instead of zero in order to facilitate the calculation of probabilities.

$$p(b, i) = \frac{f_{b,i} + s(b)}{N + \sum\limits_{b' \in A,C,G,T} s(b')} \tag{1.1}$$

The position weight matrix for a given transcription factor is finally calculated by applying equation 1.2 for each position i and base b, where $p(b)$ is the background probability.

$$W_{b,i} = log_2 \frac{p(b, i)}{p(b)} \tag{1.2}$$

The final score that assesses the likelihood for a given transcription factor to bind to a specific DNA sequence is the sum of all position specific scores of the sequences (see Equation 1.3) where l_i is the nucleotide at the position i in the sequence under study, w, the length of the sequence and S the score for the sequence.

$$S = \sum_{i=1}^{w} W_{l_i,i} \tag{1.3}$$

Equations describing the calculation of position weight matrices are taken from [Wasserman & Sandelin, 2004]. PWM binding profiles are available from databases such as JASPAR [Sandelin *et al.*, 2003, Vlieghe *et al.*, 2005], which is mainly focused on providing high quality matrices and TRANSFAC [Matys *et al.*, 2006], a more comprehensive database. Programs such as MATCH [Kel *et al.*, 2003] were developed to predict binding sites over DNA sequences using PWM. However, given the degenerate nature and short length of transcription factor binding sites, these predictions have low specificity (high false positive rates). Therefore, it is almost impossible to distinguish binding sites that are functional *in vivo* from regions with no function. For example, the matrix that describes the binding site for the myogenic transcription factor myoD predicts a total of more than 10^6 putative binding sites, whereas probably only 10^3 are functional [Wasserman & Sandelin, 2004].

1.4.2 Interspecies conservation and regulatory regions

Studies estimate that about 5% of the mammalian genome is under purifying selection [Waterston *et al.*, 2002, Roskin *et al.*, 2003]. Since coding exons comprise approximately only 1.5% of the human genome [Lander *et al.*, 2001], a fair proportion of non-coding regions of the genome is potentially functional. Among potential non-coding functional regions, *cis*-regulatory elements are good candidates [Bejerano *et al.*, 2004a]. Regulatory processes, especially the regulation of developmental processes, are particularly conserved among vertebrates. Therefore, it has been hypothesized that non-coding conserved regions were likely to be involved in regulation [Tagle *et al.*, 1988]. Long before whole vertebrate genomes were sequenced and assembled, pioneering work provided the foundation of what is now termed phylogenetic footprinting. Phylogenetic footprinting is based on the observation that functional regions accumulate fewer mutations than non-functional regions. Such studies are exemplified by the early work that focussed on the detection of *cis*-regulatory regions of the ϵ and γ globin genes [Tagle *et al.*, 1988, Gumucio *et al.*, 1992].

The publication of the complete sequence of the first human genome [Lander *et al.*, 2001, Venter *et al.*, 2001], followed by sequencing of various vertebrate genomes (from teleost fish [Aparicio *et al.*, 2002, Kasahara *et al.*, 2007] to mammals [Waterston *et al.*, 2002, Gibbs *et al.*, 2004]), combined with large scale annotation projects [Hubbard *et al.*, 2009, Curwen *et al.*, 2004, Miller *et al.*, 2007, Kent *et al.*, 2002], have established the ground for genome-wide *cis*-regulatory

region prediction. This effort was further supported by the development of sensitive DNA alignment programs [Brudno *et al.*, 2003a, Schwartz *et al.*, 2003, Brudno *et al.*, 2003b]. As an example, shuffle-LAGAN is a DNA alignment program that strategically assumes collinearity for functional elements over evolution but allows for local rearrangements [Brudno *et al.*, 2003b]. Such an approach is particularly adapted for the detection of non-coding functional elements. Several studies focusing on the regulation of specific genes of interest applied phylogenetic footprinting approaches to select putative regulatory regions prior to *in vivo* testing in model organisms. Using such an approach, regulatory regions were characterized in mouse within gene deserts next to DACH gene [Nobrega *et al.*, 2003], close to the HoxD cluster [Spitz *et al.*, 2003, Gonzalez *et al.*, 2007] as well as the gene desert adjacent to the iroquois gene cluster (in transgenic xenopus and zebrafish embryos) [de la Calle-Mustienes *et al.*, 2005]. Various genome-wide analyses developed different purely computational procedures to predict putative vertebrate *cis*-regulatory regions by detecting parts of the genome under purifying selection [Dermitzakis *et al.*, 2002, Bejerano *et al.*, 2004a, Elnitski *et al.*, 2003]. *In silico* detection of putative *cis*-regulatory regions combined with efficient transgenic assays led to large-scale programs for the characterization of long-range *cis*-regulatory regions in the human genome [Pennacchio *et al.*, 2006]. In this study, they assessed 167 putative enhancers that can be accessed at the VISTA Enhancer Browser (http://enhancer.lbl.gov).

The detection of regions under purifying selection is usually undertaken by aligning genomes of various evolutionary distances. The evolutionary depth

in terms of species surveyed has an impact on the type of regions that will be detected. For example, sequence comparisons that involve human and teleost fish are highly specific in detecting functional sequences but most likely will be limited to the discovery of a small subset of regulatory regions related to specific biological processes such as development. Non-coding conserved regions are indeed enriched for developmental regulators [Sandelin *et al.*, 2004, Woolfe *et al.*, 2005, Plessy *et al.*, 2005], but many regulatory regions are clade or species specific and would be missed by using conservation between distant genomes.

To reduce the bias induced by conservation, methods involving the comparison of closely related species, such as phylogenetic shadowing [Boffelli *et al.*, 2003], have been developed. Phylogenetic shadowing, which was applied to primate sequences is encouraging but lacks resolution and is not precise enough to detect accurate boundaries of *cis*-regulatory regions. The sequence divergence between closely related primate sequences is indeed too low to allow proper detection of regions under purifying selection, unless a very large number of sequences is used.

Conservation has proven to be a powerful tool to detect functional regions but this approach is not helpful to assign a specific function to the region of interest. Moreover, not all conserved regions seem to be functional. In an experiment conducted by Nobrega *et al.*, the deletion of conserved non-coding regions from a mouse showed no measurable phenotypic effect and produced a perfectly viable mouse [Nóbrega *et al.*, 2004]. Finally, based on the comparison of experimentally characterized binding sites between human and mouse, Odom *et al.* demonstrated that about two third of binding sites do not align between human and mouse

genomes [Odom *et al.*, 2007]. So to increase the sensitivity of predictions, phylogenetic footprinting is more often used in conjunction with other prediction methods.

Interspecies conservation is a powerful filter to reduce the false positive rate from binding site predictions [Lenhard *et al.*, 2003] (discussed in the previous section). This approach is now used by various binding site predictors such as Consite [Lenhard *et al.*, 2003] and rVista [Loots *et al.*, 2002].

1.4.3 *Cis*-regulatory module predictions

As described in Section 1.2, functional transcription factor binding sites are typically organized into clusters. A cluster, also called a *cis*-regulatory module (CRM), is basically defined as a region where an overrepresentation of binding sites for a few transcription factors is observed. New algorithmic approaches such as hidden markov models [Frith *et al.*, 2001, Frith *et al.*, 2003] or statistical methods [Schones *et al.*, 2007, Johansson *et al.*, 2003, Rajewsky *et al.*, 2002] were dedicated to this problem. Currently, many applications are available, such as Cluster-Buster [Frith *et al.*, 2003], MSCAN [Johansson *et al.*, 2003] and Ahab [Rajewsky *et al.*, 2002].

Cross-species comparisons can improve CRM predictions [Sinha *et al.*, 2004] and have been integrated by programs in the CREME framework [Sharan *et al.*, 2003] or module finder [Philippakis *et al.*, 2005]. Using such combined approach, Blanchette *et al.* predicted 118,000 CRMs from which a fraction was confirmed by ChIP-CHIP analysis [Blanchette *et al.*, 2006].

1.4.4 Detection of regulatory regions common to co-expressed genes

Co-expressed genes are genes that show same expression pattern under specific conditions and in the same tissues, therefore they are more likely to be under the regulation of identical TFs [Chang *et al.*, 2006]. Expression data are a useful resource for predicting binding sites or CRMs which may take part in shared regulatory processes. With the advent of microarray gene expression technologies, large-scale expression data for different cell types, diseases or conditions, are now available. In the context of developmental regulation, high throughput *in-situ* hybridization data for model organisms such as zebrafish [Sprague *et al.*, 2006] are also an valuable source of information.

New approaches and algorithms were developed to predict over-represented binding sites in the promoters of co-regulated genes [Chang *et al.*, 2006]. The use of co-expression data was also applied to the identification of CRMs, for example, muscle specific CRMs were predicted in the vicinity of C. *elegans* genes expressed in muscle [Zhao *et al.*, 2007]. The co-regulation approach was also combined with phylogenetic footprinting to improve predictions quality (for example [Grad *et al.*, 2004] in drosophila blastoderm development).

1.5 *In vivo* characterization of *cis*-regulatory regions

Regardless of the method used to predict *cis*-regulatory regions, such as comparative genomics or more advanced methods, tools are needed to characterize the regulatory specificity of a *cis*-regulatory region. Although cell culture is a powerful system to learn about enhancers, *in vivo* characterization is needed for proper spatio-temporal characterization of a putative regulatory region.

Therefore, various biological frameworks with biomolecular techniques and model organisms were specifically developed to properly monitor the expression of putative regulatory regions. Two main elements are important when developing a biomolecular framework to characterize long-range regulators: i/ the model organisms, and ii/ the reporter construct.

1.5.1 Vertebrate model organisms to characterize long-range regulation

Model organisms are chosen depending on the goal of the study, its scale (e.g. the number of regions to be tested) and the technical abilities of the laboratory. Among the different model organisms available to test *cis*-regulatory regions, mouse is one of the most commonly used. Mouse, as model organism has many advantages: i/ It is relatively close to human (it diverged approximately 60 million years ago with human), which is valuable when studying human diseases. ii/ It has a short gestation time ranging from 18 to 21 days. iii/ The mouse full genome sequence is available [Waterston *et al.*, 2002] providing a genome-wide base for the search of regulatory regions. However, in the context of regulatory region characterization, mouse has drawbacks: i/ In order to perform whole mount stained (fixation of the embryo prior to imaging), the embryo is killed, which limits the observation of expression pattern to only one developmental stage (e.g. [Pennacchio *et al.*, 2006]). ii/ Mouse genetic experiments are expensive and time consuming, which does not make them scalable to large experiments such as large-scale enhancer characterization.

Although more evolutionary distant from human, teleost fish are appropriate model organisms for developmental genetics and human diseases [Ishikawa, 2000,

Zon, 1999]. The main advantage of teleost fish are: i/ Their short generation time (medaka hatching stage is reached in 9 days). ii/ Their transparent ex-utero embryos allows live observation of the developing fish, therefore, contrasting with mammalian model organisms, expression patterns can be observed at different developmental stages. iii/ Their relatively compact genome makes *in silico* detection of *cis*-regulatory region more convenient.

Among various teleost fishes, the *Takifugu rubripes*, Fugu, with a compact genome of about 400 Mb [Ishikawa, 2000] is a natural candidate for the detection of *cis*-regulatory regions. However, this organism is not currently bred under laboratory conditions [Müller *et al.*, 2002]. Zebrafish, although not as compact (genome size of 1700 Mb [Ishikawa, 2000]), is a common laboratory animal and has been previously used to functionally characterize promoters [Amacher, 1999, Westerfield *et al.*, 1992, Müller *et al.*, 1999]. Finally, medaka, although not used as often as other model organisms to characterize regulatory regions, combines the advantages of a small genome size (about 800 Mb [Ishikawa, 2000]), small generation time and ease of breeding in laboratory. Its suitability to functionally characterize promoters with a green fluorescent protein reporter was shown in previous studies, such as in [Tanaka *et al.*, 2001]. On a final note, there are other vertebrate model organisms that have not been mentioned here such as the frog *Xenopus tropicalis*, which is also becoming a popular organism for testing *cis*-regulatory regions (reviewed in [Khokha & Loots, 2005]).

1.5.2 Reporter constructs to characterize *cis*-regulatory regions

To be able to observe the regulatory specificity of a specific genomic region, the DNA corresponding to this region needs to be inserted into the model organism. Two components are necessary: the vector (the carrier of the construct) and a reporter gene (gene that would be expressed if the region tested has regulatory properties). In the case of long-range enhancer testing, a minimal promoter is added to the construct. This minimal promoter can either be the promoter of the gene suspected to be the target of the tested enhancer (endogenous minimal promoter) or be one generic promoter that is common to all tested regions.

The injection of plasmid DNA or artificial chromosome in one cell embryo is the most common transgenesis method to test regulatory regions. When large constructs need to be tested, artificial chromosomes such as bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs) (reviewed in [Giraldo & Montoliu, 2001]) are particularly adapted but their efficiency to produce germ-lines is low [Amsterdam & Becker, 2005]. Plasmid are much smaller but convenient when regions to be tested are small.

The reporter gene is a gene that is cloned in the vector and reports the expression driven by the element(s) of the construct with regulatory specificity (e.g. the enhancer). This gene must code for a protein that is visible under certain filters so the specificity of the regulatory region can be reported at the appropriate time and in the appropriate tissue. LacZ has been successfully used in mouse [McFadden *et al.*, 2000, Bagheri-Fam *et al.*, 2006, Gonzalez *et al.*, 2007]

but is not popular in teleost as it is frequently inactivated after germ-line passage in zebrafish [Amsterdam & Becker, 2005]. The green fluorescent protein (GFP) (reviewed in [Tsien, 1998]) is widely used as a reporter gene in teleost. In combination with new developments in microscopy, GFP is suitable to identify expression within single cells [Megason & Fraser, 2003] and therefore has been successfully used to charaterize enhancer activity GFP in live xenopus and zebrafish [de la Calle-Mustienes *et al.*, 2005]. An example of basic steps to follow for enhancer *in vivo* testing is presented Figure 1–2.

Alternatively, cloning can be avoided by co-injecting the putative enhancer with a reporter construct in the embryo. This method was developed in fish by Muller *et al.* [Müller *et al.*, 1999] to detect an enhancer element of sonic hedgehog in zebrafish. Co-injection (reviewed in [Müller *et al.*, 2002]) was more recently applied in various studies to test enhancer elements [Rastegar, 2002, Woolfe *et al.*, 2005, Sanges *et al.*, 2006]. Although this method is time and cost effective, results are difficult to interpret. Expression is observed in a transient manner which means that since the construct does not integrate into the genome, some cells carry the construct whereas other do not. Expression pattern is therefore mosaic and to be meaningful, data need to be compiled from many embryos. Such compilation becomes particularly difficult if different times of development are under study [Gómez-Skarmeta *et al.*, 2006].

There have been some attempts to generate reporter gene expression in selected structures by randomly inserting in the genome of a model organism a construct composed of a minimal promoter coupled with a reporter gene

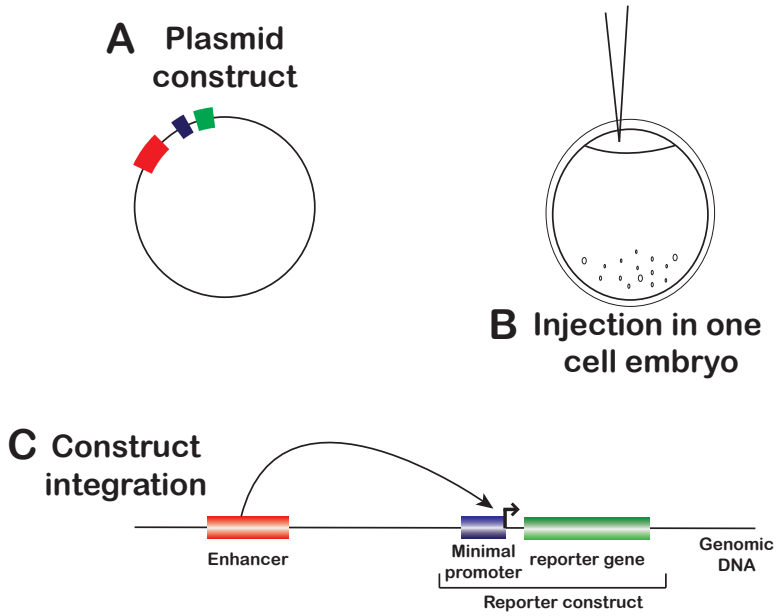

Figure 1–2: **Major basic steps for *in vivo* testing of enhancer activity with a plasmid construct.** A/ A construct is designed (here a plasmid) with the region to be tested (red), a minimal promoter (blue) and a reporter gene (green). B/ After plasmid growth and purification, the construct is injected in one cell embryo with restriction enzymes. C/ The linearized DNA integrates the genome and if the region tested has regulatory property, the reporter gene is activated.

(enhancer trapping). In some cases the construct, whose location of insertion in the genome is retrieved by PCR, will be activated by nearby regulatory elements resulting in the selective activation of the reporter gene [Parinov *et al.*, 2004, Choo *et al.*, 2006]. But this method can not be used to target specific regions and interesting outcomes are tied to the random insertion of the construct in a region of interest.

1.6 Genetic instability and long-range regulation

Gross chromosomal rearrangements are an alteration of the genetic linkage of two DNA fragments due to recombination of DNA fragments [Aguilera & Gómez-González, 2008]. Rearrangements are potentially detrimental since they may result in loss of genetic material or transposition of DNA into a different genetic environment. Therefore, rearrangements are only fixed in regions of the genome where they are not deleterious (or confer an advantage) and consequently their localization reflects the variations of functional pressure on the genome. In this thesis, we study evolutionary breakpoints in the context of long-range regulation. In order to introduce this work, we discuss in this section: i/ the most common types of rearrangements, ii/ examples of the main mechanisms causing these rearrangements, iii/ the distribution of evolutionary breakpoints, and iv/ the link between evolutionary rearrangements and long-range regulation.

1.6.1 Gross rearrangements

Deletions, inversions, duplications and translocations are among the common types of gross chromosomal rearrangements [Lupski & Stankiewicz, 2005,

Gu *et al.*, 2008]. Gross rearrangements are a common cause of genetic variation and can affect anywhere from a part of a gene to millions of base pairs [Coghlan *et al.*, 2005, Lupski & Stankiewicz, 2005]. Descriptions of gross rearrangements are based on [de Sá, 2007] and [Lodish *et al.*, 1995], p 267.

A deletion is the loss of a segment of DNA (Figure 1–3, A). If this loss span the centromere, the rearranged chromosome is called acentric; acentric chromosomes are lost during meiois. Deletion can have strong phenotypic effect if the deleted region contains genes or regulatory regions.

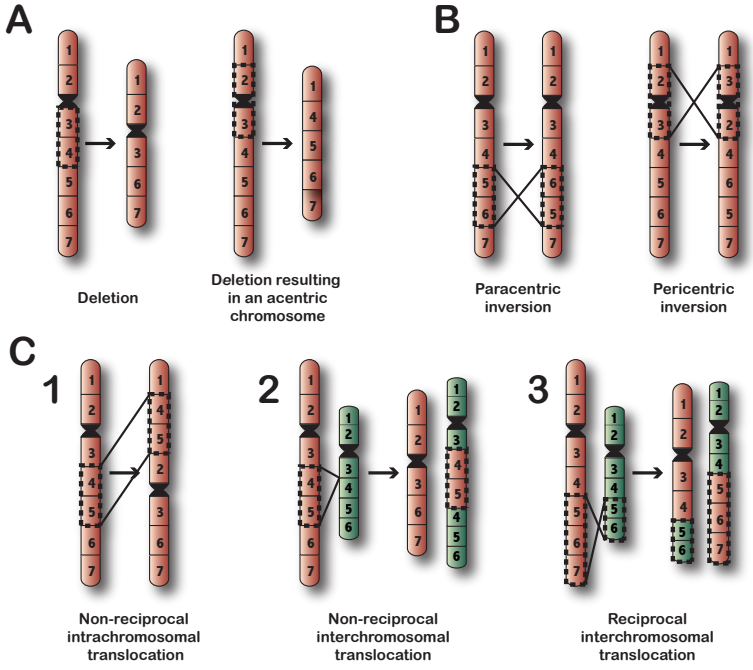

Figure 1–3: **Gross rearrangements**. Major gross rearrangements represented by A/ deletion, B/ inversion, and C/ translocation. This figure is particularly inspired from [de Sá, 2007].

An inversion occurs when a segment of DNA is broken and then rejoigned in the wrong orientation. Two types of inversions exist: i/ inversions located excusively on one chromosome arm called paracentric inversions, and ii/ pericentric inversions when inversion spans the centromere (Figure 1–3, B). Although there is no loss of genetic material, if the breakpoints disrupt a gene, a regulatory region or break the physical association between a gene and a regulatory region, there may be phenotypic consequences.

A translocation is the relocation of a chromosome segment to a different part of the genome. Different translocation events are: i/ non-reciprocal intrachromosomal translocation, where a segment of DNA is moved to another region of the same chromosome (Figure 1–3, C1), ii/ non-reciprocal interchromosomal translocation, where a segment of DNA is moved to another chromosome (Figure 1–3, C2), and iii/ reciprocal interchromosomal translocation, where DNA segments are exchanged between two chromosomes (Figure 1–3, C3). All types of translocation involve physical breakage of the chromosome and so have the possibility to disrupt a gene, regulatory region or to disrupt the regulation of a gene.

Among other rearrangements, **duplication** is when a part of the chromosome is identically reproduced (such as in segmental duplication) and an **insertion** corresponds to the addition of a segment of DNA within the chromosome.

1.6.2 Mechanisms responsible for evolutionary rearrangements

Two main mechanisms are currently proposed to explain gross rearrangements: non-allelic homologous recombination (NAHR) and non-homologous end joining (NHEJ).

NAHR is associated with recurrent rearrangements (rearrangements with similar lengths, observed in many individuals and responsible for clustered breakpoints [Gu *et al.*, 2008]) and is the most common mechanism responsible for large structural variations within a population [Kidd *et al.*, 2008]. NAHR mediated gross rearrangements have been linked to specific genomic structures such as transposons, minisatellites, triplet repeats, and LCR [Shaw & Lupski, 2004, Gu *et al.*, 2008]. It appears that in meiosis or mitosis, the misalignment of two non-allelic copies of regions of high identity (such as LCRs) can mediate a homologous recombination that result in gross rearrangements [Gu *et al.*, 2008]. Depending on the location and orientation of these repeats, NAHR results in an inversion, deletion, duplication or translocation [Gu *et al.*, 2008, Aguilera & Gómez-González, 2008].

NHEJ [Roth *et al.*, 1985, Roth & Wilson, 1986] is another eukaryotic mechanism involved in double strand break repair. In the context of gross rearrangements, this mechanism is mainly associated with non-recurrent rearrangements. Basically NHEJ process involves the detection of double strand break, modification of the ends and ligation [Gu *et al.*, 2008]. Therefore, NHEJ does not require specific repeat sequences but may be stimulated by local genomic architecture such as the presence of repeat elements [Shaw & Lupski, 2004, Stankiewicz *et al.*, 2003]. For example, NHEJ have shown microhomology at their junctions [Nobile *et al.*, 2002, Toffolatti *et al.*, 2002]. They also may be found in regions that are more prone to double strand breaks such as highly transcribed regions and non-B DNA confomation. NHEJ may

be involved in reciprocal translocations, deletions, inversions or insertions [Shaw & Lupski, 2004, Aguilera & Gómez-González, 2008].

NAHR and NHEJ are both involved in double strand break repair, therefore are tightly linked to replication stress. Non-B DNA conformation that may favor double strand breaks [Wells, 2007] is also believed to be a cause for rearrangements [Bacolla *et al.*, 2004, Wells, 2007].

1.6.3 Chromosome rearrangements are not random

If not detrimental to the fitness of the individual (e.g. it does not disrupt a gene or a regulatory region), a given rearrangement can become fixed in evolution. By comparing genomes from different species these rearrangements can be detected.

The distribution of rearrangements in vertebrate evolution has been until recently a controversial subject. A milestone paper paper published in 1984 by Nadeau and Taylor [Nadeau & Taylor, 1984] influenced the field of chromosome evolution for a few decades claiming that chromosomal rearrangements are randomly distributed along the genome. To do so, they first introduced the concept of synteny blocks, defined by the conservation of marker order (such as genes) between two (or more) genomes. They worked on limited dataset (in 1984 no full genome sequence was available) and had to estimate the distribution of synteny block lengths to calculate the number of breakpoints [Pevzner & Tesler, 2003]. Their estimation was unfortunately wrong and consequently their conclusions on breakpoint distribution have not held true.

However this study stimulated the next wave of studies on rearrangements. Pevzner and Tesler [Pevzner & Tesler, 2003, Peng *et al.*, 2006] demonstrated that evolutionary breakpoints accumulate in certain regions in mammalian evolution. They called such accumulation of breakpoints in "fragile regions", "breakpoint reuse". Those studies did not focus on the functional explanation of such breakpoint reuse but laid new perspectives on mammalian chromosome evolution. The breakpoint reuse model is now widely accepted and confirmed by other studies [Larkin *et al.*, 2003, Murphy *et al.*, 2005]. However the causes and mechanisms responsible for the fact that some regions are more often rearranged than others are not yet totally elucidated.

1.6.4 Long-range regulation may prevent evolutionary rearrangements

The probability of fixation of rearrangement in a population depends on the fitness of the affected individual. It has long been assumed that breakpoints outside of genes are unlikely to be deleterious. There is now evidence that this is not the case and that long-range regulation plays a major role in restricting the space of possible rearrangements. We review in this section the different studies that associated evolutionary breakpoints with long-range regulation.

First, non-coding conserved regions and genes are not homogeneously distributed along the human genome and about 25% of the genome is composed of gene deserts (genomic regions without protein coding genes) [Venter *et al.*, 2001].

Those gene deserts are enriched for non-coding conserved regions that are enriched for regulatory functions and are rarely broken by evolutionary breakpoints [Ovcharenko *et al.*, 2005]. Moreover, these non-coding conserved regions are maintained in synteny with neighboring genes [Nobrega *et al.*, 2003, de la Calle-Mustienes *et al.*, 2005, Woolfe *et al.*, 2005] (see Section 1.4 for more references). These genes, enriched for transcriptional and developmental biological processes are usually maintained in synteny throughout vertebrate evolution, suggesting that a link between non-coding conserved regions and these genes is crucial and cannot be broken.

Some regions of the genome such as the Irx or HoxD clusters are particularly conserved among vertebrates (for both regulatory and coding regions). The HoxD cluster shows unusual conservation where both genes and enhancers's synteny is kept among most vertebrates. Spitz *et al.* [Spitz *et al.*, 2003] established that within this region - organized in what they termed a global control regions (GCR) - a single enhancer can regulate many genes. These multiple functional interactions between regulatory regions and genes may explain and reinforce the observed conservation of synteny. Later studies reinforced the idea that such conserved synteny at the Hox locus may be caused by such regulatory constraints [Lee *et al.*, 2006]. The range of action of *cis*-regulatory regions is an additional element to take into account when dealing with constraint imposed by such regions. Some regulatory regions, such as an enhancer regulating Shh gene are located as far as 1 Mb from the target gene [Lettice *et al.*, 2003], consequently imposing a functional constraint over a long genomic region.

Those observations taken together led to a model where syntenic relationship between regulatory regions and their putative genes are maintained in synteny throughout evolution by functional constraints [Mackenzie *et al.*, 2004, Becker & Lenhard, 2007, Ahituv *et al.*, 2005, Goode *et al.*, 2005]. The rationale behind this model is that a breakage of the functional link between a regulatory region and the target gene may result in the mis-regulation of the gene and phenotypic change. Such changes may impact the fitness of the mutated individual, especially if the gene is involved in fundamental function such as transcriptional or developmental processes. Therefore, evolutionary rearrangements are potentially a powerful tool to be used in long-range regulation studies.

1.7 Thesis outline and hypotheses

Disruption of long-range regulatory regions is the cause for various developmental genetic diseases such as Williams-Beuren syndrome [Merla *et al.*, 2006], aniridia [Kleinjan *et al.*, 2001], Pierre Robin syndrome [Benko *et al.*, 2009] or Campomelic dysplasia [Pop *et al.*, 2004]. In these diseases, the transcriptional control of key developmental genes is disrupted, which leads to severe phenotypic effects. In some cases, *cis*-regulatory regions whose action are disrupted are located over a few hundred of kb from the gene that is responsible for the phenotype (position-effect related diseases are reviewed in [Kleinjan & van Heyningen, 2005]). Proper identification of *cis*-regulatory regions involved in developmental regulation remains a key element to better understand the cause of many genetic diseases.

Evolutionary rearrangements occur non-randomly on the human genome [Pevzner & Tesler, 2003]. The particular predilection of certain regions to undergo

and fix evolutionary breakpoint is probably influenced by local functional pressure. Long-range regulatory regions are involved in the regulation of genes as far as 1 Mb and the disruption of the association between a gene and a *cis*-regulatory region may result in strong phenotypic effects. We hypothesized that long-range regulation may be a major force in maintaining the genome integrity and therefore prevent evolutionary breakpoints from occurring in certain genomic regions. To identify the forces at work in shaping the human genome, we present a machine-learning method in **Chapter 2** to characterize local predisposition for evolutionary breakpoints on the human genome.

Various methods exist to identify putative *cis*-regulatory regions on vertebrate genomes, however few studies attempt to predict which genes they regulate. Such a prediction dataset would be useful to the scientific community, particularly in the context genetic diseases whose cause is the mis-regulation of particular genes, to prioritize *in vivo* testing of putative regulatory regions. The objective of the work, presented in **Chapter 3**, is therefore to develop a method to predict putative target genes for human *cis*-regulatory regions. This work based on observations made in Chapter 2 establishes a new method to predict which genes a *cis*-regulatory region may regulate based on evolutionary rearrangements. We present a new expectation-maximization algorithm based on evolutionary rearrangements between human and 16 vertebrate genomes to calculate the likelihood of functional associations between pairs of genes and regulatory regions.

The proper understanding of long-range regulation in the context of embryonic development depends on proper *in vivo* characterization of predicted

regulatory regions. *In vivo* characterization not only confirms that the region is indeed functional, but also allows a precise monitoring of spatio-temporal expression patterns in development. However, few methods exist to characterize regulatory regions in living model organisms in a fast and reliable manner. In **Chapter 4**, we present an original combination of *in silico* predictions of regulatory regions and a novel experimental method that tests those predictions in medaka fish.

Finally, in **Chapter 5** we discuss the role of long-range regulation in shaping the human genome throughout evolution, as well as how this may impact the prevalence of disease related rearrangements.

CHAPTER 2
Long-range regulation is a major driving force in maintaining genome integrity

2.1 Preface

Evolutionary rearrangements are rare in various regions of the human genome such as the Hox clusters. Such regions are usually involved in developmental processes where genes are known to be under tight regulatory control. The hypothesis behind this project is that long-range regulatory regions may be a major force in preventing evolutionary breakpoints to occur and become fixed in a population in some regions of the genome as such events may be too deleterious. Our approach consists of i/ developing a novel method to predict the different levels of susceptibility for evolutionary breakpoints on the human genome, and ii/ analyzing those different regions to determine if long-range regulation is related to rearrangements.

The method to predict evolutionary breakpoints we developed is based on a linear regression classifier using features related to regulatory regions and genes. The main advantage of this approach compared to the direct study of evolutionary breakpoints is that the impact of each feature on the prediction can be assessed. We show a clear functional dichotomy between breakpoint-susceptible regions and breakpoint-refractory regions that confirms our hypothesis that long-range

regulation impacts the fixation of evolutionary breakpoints. This investigation provides new insight on long-range regulation and its evolutionary effects. This Chapter was published in BMC Evolutionary Biology as follows:

- **Emmanuel Mongin**, Ken Dewar, Mathieu Blanchette. Long-range regulation is a major driving force in maintaining genome integrity. *BMC Evol Biol*, 9(1):203, Aug 2009.

2.2 Abstract

Background: The availability of newly sequenced vertebrate genomes, along with more efficient and accurate alignment algorithms, have enabled the expansion of the field of comparative genomics. Large-scale genome rearrangement events modify the order of genes and non-coding conserved regions on chromosomes. While certain large genomic regions have remained intact over much of vertebrate evolution, others appear to be hotspots for genomic breakpoints. The cause of the non-uniformity of breakpoints that occurred during vertebrate evolution is poorly understood.

Results: We describe a machine learning method to distinguish genomic regions where breakpoints would be expected to have deleterious effects (called breakpoint-refractory regions) from those where they are expected to be neutral (called breakpoint-susceptible regions). Our predictor is trained using breakpoints that took place along the human lineage since the metatheria divergence. Based on our predictions, refractory and susceptible regions have very distinctive features. Refractory regions are significantly enriched for conserved non-coding elements as

well as for genes involved in development, whereas susceptible regions are enriched for housekeeping genes, likely to have simpler transcriptional regulation.

Conclusions: We postulate that long-range transcriptional regulation strongly influences chromosome break fixation. In many regions, the fitness cost of altering the spatial association between long-range regulatory regions and their target genes may be so high that rearrangements are deleterious. Consequently, only a limited, identifiable fraction of the genome is susceptible to genome rearrangements.

2.3 Background

Genomes evolve through a series of local mutations as well as larger-scale genome rearrangements (such as inversions, translocations and duplications) where one or more chromosomes break in one or more locations (called breakpoints) and fragments are reorganized. Just as for point mutations, the likelihood that a particular rearrangement becomes fixed in the population depends (in part) on the fitness of the mutated individual [Coghlan *et al.*, 2005]. In comparative genomics, the comparison of gene orders in different species (i.e. of those rearrangements that have become fixed in their respective populations) sheds light on genome evolution [Pevzner & Tesler, 2003, Peng *et al.*, 2006] and phylogenetics [Sankoff & Nadeau, 2003, Blanchette *et al.*, 1999, Yue *et al.*, 2008].

In 1984, Nadeau and Taylor published a paper where breakpoints of genome rearrangements (chiefly inversions and translocations) between human and mouse were modeled as occurring randomly and uniformly in the genome [Nadeau & Taylor, 1984], a hypothesis later supported by Sankoff *and* Trinh

[Sankoff & Trinh, 2005]. This model relied on the implicit assumption that most breaks of synteny (disruption of the order of markers, genes or regulatory elements, along a chromosome caused by genome rearrangements) do not have significant functional implications. However the availability of more genomes to undertake comparative genomic studies and new algorithms to identify breakpoints increased both the resolution and the completeness of the analysis. This led to a model where rearrangement breaks do not occur uniformly but instead where some regions, termed 'evolutionary hotspots', are more prone to breakage, resulting in a high level of breakpoint reuse [Pevzner & Tesler, 2003, Peng *et al.*, 2006]. Although it is now generally accepted that evolutionary breakpoints (i.e. rearrangement breakpoints that became fixed in a particular population) are not uniformly distributed on the human genome, the reasons why some regions tend to fix chromosomal rearrangements more than others still remains unclear and to date, no satisfactory explanation has yet been given at the whole genome level.

Long-range regulation has been hypothesized to be one of the elements that favor conservation of synteny in certain regions of the genome [Mackenzie *et al.*, 2004]. Studies focusing on specific vertebrate regions, such as the Hox cluster [Lee *et al.*, 2006] or the Shh locus [Goode *et al.*, 2005], where a strong selective pressure is obviously at work, illustrated the notion that regulatory regions surrounding those loci could induce evolutionary constraints that maintain the integrity of the genome. Kikuta *et al.* [Kikuta *et al.*, 2007] and Engstrom *et al.* [Engström *et al.*, 2007] established that some regions are under the influence of what they designated as genomic regulatory blocks (GRBs), which control the expression of developmental genes

over a large genomic region, and showed that the synteny around those GRBs is maintained. To date no genome wide analysis has been undertaken to uncover the different levels of susceptibility of the human genome to breakpoints. Such information is crucial to better understand the forces preventing breakpoints from being fixed in evolution.

In this paper, we propose a new approach to estimate the susceptibility of regions of the human genome to tolerate breakpoints. Our method is trained to recognize these regions based on the presence of coding, conserved non-coding elements (assumed to be enriched for regulatory regions), and their putative interactions. We were able to define two types of regions: those that are prone to accept evolutionary breakpoints and those that are refractory to breakpoints. The analysis of those regions uncovers features that shed some light on the underlying mechanisms of selection against rearrangements. This suggests that long-range regulation is a major driving force in maintaining genome integrity.

2.4 Results

2.4.1 Synteny mapping

In this analysis, we study breakpoints that occurred along the human lineage since the metatheria divergence (eutherians vs marsupials split). These breakpoints can be identified through the comparison of the human genome to that of a marsupial (here, opossum [Mikkelsen *et al.*, 2007b]), and an outgroup, chicken [Hillier *et al.*, 2004]. Identifying breakpoints requires the detection of unique, conserved markers, present in each of the species studied.

Based on whole genome 'liftover chains' pairwise alignments [Kent *et al.*, 2002] (human/opossum and human/chicken) which are a hierarchical collection of sequences of gapless aligned blocks, we mapped human markers to opossum and chicken. Markers are of two types; (i) non-coding conserved regions which are considered enriched for regulatory elements and (ii) coding regions. We identified 116 331 markers, each present exactly once in human, opossum, and chicken. We call these *amniote* markers. We also identified 93 802 *metatherian* markers, conserved between human and opossum, but absent in chicken. Metatherian markers will not be used to define breakpoints (because no outgroup is available to determine their ancestral status), but will later be taken into consideration in our prediction of break-prone regions.

A breakpoint between human and opossum (resp. chicken) is defined as a pair of amniote markers that are adjacent in human but not opossum (resp. chicken). Because establishing the orthology of human, opossum, and chicken markers is error-prone, we deliberately removed from further consideration 383 markers that are flanked by breakpoints on both sides in either the human/opossum or human/chicken comparisons. Although some of these breakpoints may be real, we argue that most of them are likely due to incorrect genome assembly or orthology mapping. This reduced set of markers was then used to define a set of 845 reliable human/opossum and 1546 human/chicken breakpoints. The intersection of the human/opossum and human/chicken breakpoints, which corresponds to 412 breakpoints that took place along the human lineage since the divergence of

metatherians, is called the set of *human breakpoints* and is the focus of our study in the rest of this paper.

As expected, breakpoints within protein-coding genes are rare, forming only 2.7% of human/opossum breakpoints and 3.4% of human/chicken breakpoints. Many of these intragenic breakpoints are likely to be the result of incorrect gene annotation. For example, annotated genes such as MPP4 are made of multiple spliced variants gathered together in one gene, but it could very well correspond to two independent transcriptional units. Because of this, the few human breakpoints occurring within annotated genes were assigned to the left side of the gene. Annotated genes now being free of breakpoints, all markers within them (whether they are coding or not) were collapsed into a single meta-marker called a coding marker. The number of markers of each type is given in Table 2–1.

Table 2–1: **Number of markers for each type and conservation level.** Coding markers are genes as annotated in EnsEMBL and non-coding markers are non-coding conserved regions (taken from the UCSC 28-way alignment). A human marker (coding or non-coding) conserved only at between human and opossum is labeled as metatheria marker. A human marker conserved both with opossum and chicken is labelled amniote marker.

Marker type	metatheria	amniote	total
coding	4335	9951	14286
non-coding	46914	26217	73131

What factors determine the likelihood that a particular rearrangement becomes fixed in a population? The main factor is likely to be the difference of fitness between individuals with and without the rearrangement. While a given rearrangement is rarely going to be beneficial to the affected individual, it may very

well be detrimental. Three situations may be particularly deleterious: (i) when a breakpoint occurs within a gene, (ii) when a breakpoint occurs between a gene and a *cis*-regulatory element for that gene, thus separating this gene from its regulator, and (iii) when a rearrangement brings a regulatory element in the vicinity of a gene leading to its mis-regulation. See also [Kirkpatrick & Barton, 2006] for population genetics considerations. To predict the potential effect of a breakpoint at a given genomic position, it is useful to look at the context within which the breakpoint happened (the ancestral state), rather than the result of that rearrangement (the derived state). Of course, we do not have access to the exact ancestral state surrounding each breakpoint. However, because breakpoints are rare and, for the most part, separated by fairly large genomic distances [Pevzner & Tesler, 2003], the genome of the closest extant species outside the lineage on which the rearrangement occurred provides a good approximation of that local ancestral state. In our case, since we focus on breakpoints on the human lineage, the ancestral state can be approximated using the opossum local context. Moreover, we only consider syntenic blocks consisting of at least of two markers, which excludes most micro rearrangements. This approximation does not take in account events that may have occurred in the same region on the opossum branch after divergence, but since the goal of this study does not require a high level of precision, this method is, in our point of view, sufficient. Moreover, trying to computationally infer the real ancestral state could lead to errors that may add noise and not improve the prediction.

Figure 2–1 shows that there is a strong enrichment for breakpoints occurring between two coding markers and a strong depletion of breakpoints flanked by one or two non-coding markers, supporting the hypothesis that regulator/target-gene relations severely constrain breakpoint fixation. Following this observation, we aimed at understanding better what properties of a given genomic region increases or decreases its likelihood of being involved in a breakpoint that would become fixed in the population, and to train a classifier to predict break-prone inter-marker regions based on their context. Our data set thus consisted of 383 positive examples (the human-lineage breakpoints, considered in their ancestral (opossum) context), and 35 586 negative examples (inter-marker regions without breakpoints). It should be noted that inter-marker regions over 1Mb in opossum and human have been removed from consideration because breakpoints couldn't be located sufficiently precisely.

2.4.2 Features used for breakpoint prediction

Two types of features were used for breakpoint prediction (see Figure 2–2); the local density of functional elements and the association between non-coding putative regulatory regions and genes. The local density of each type of functional elements (coding-metatherian, coding-aminote, noncoding-metatherian, noncoding-amniote) is measured as a weighted count of such elements in a 2Mb-window centered on the region of interest. The weight of an element decreases as a function of its distance from the center of the window, as $w(d) = 1/\log(d)^{\alpha}$. Choosing $\alpha = 0$ gives the same weight to all elements within the window, whereas a high α factor ($\alpha = 3$) gives a much higher weight to elements close to the center.

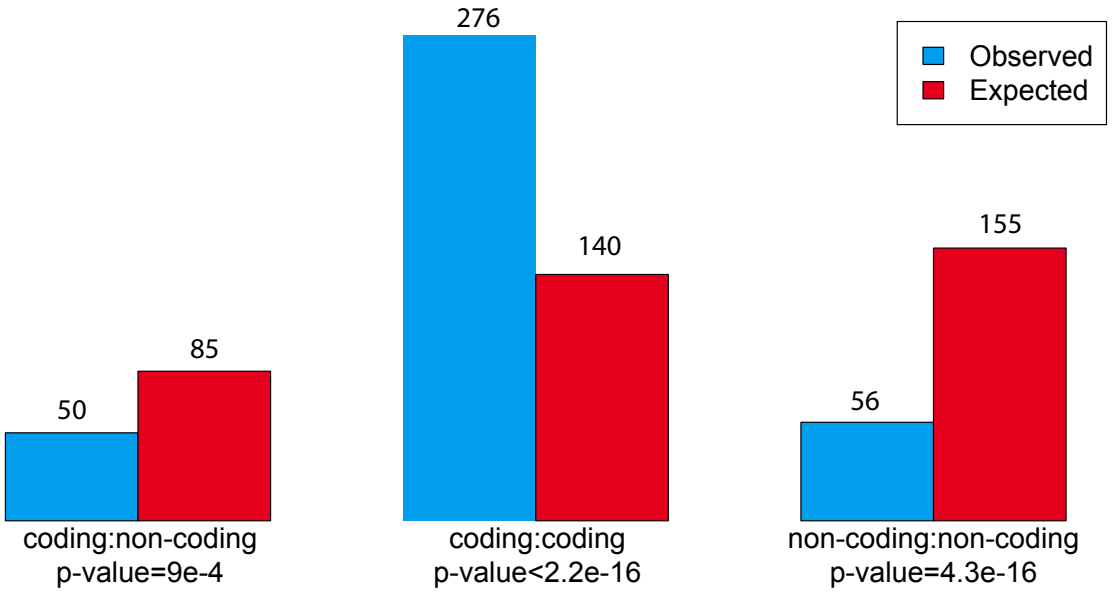

Figure 2–1: **Breakpoints and surrounding markers.** Observed (blue) and expected (red) number of breakpoints depending on the types of flanking ancestral markers. The expected number of breakpoints was calculated based on the total size of inter-marker regions of each type. P-values were calculated with a chi-square test.

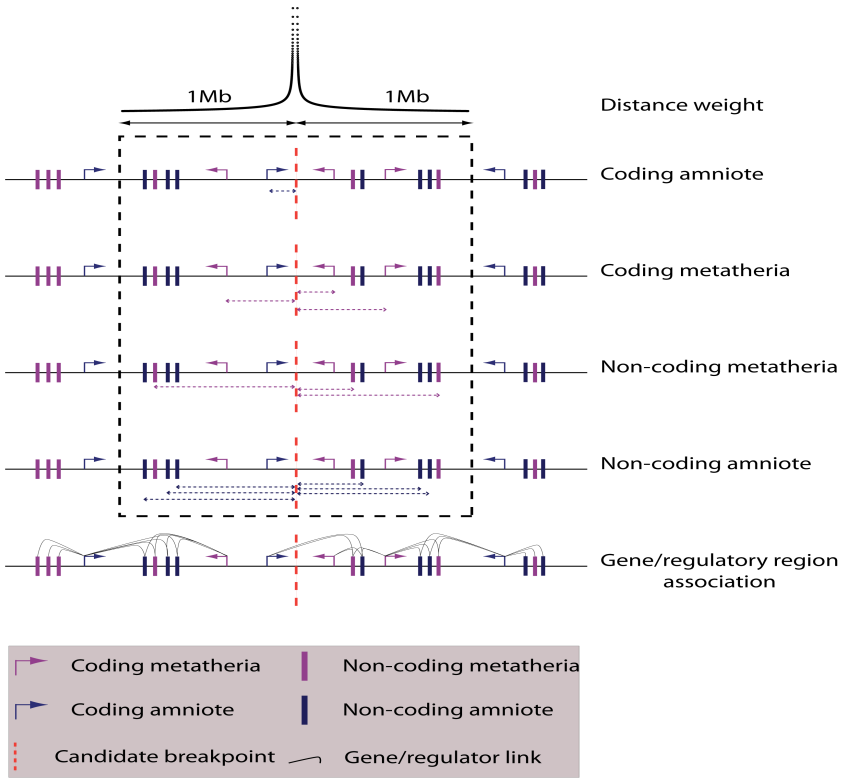

Figure 2–2: **Feature selections.** The diagram represents the different types of features used by our predictor. For the local density of functional elements, the categories are coding amniote, non-coding amniote, coding metatheria, non-coding metatheria within a 2Mb window. For each of these categories, the distance from the candidate breakpoint to the center of inter-marker region considered is taken into account and various weighting factors termed α are applied. The effect of an α factor of 2 is presented at the top of the figure. In that case markers close to the breakpoint have a high weight whereas distant markers only bring a small contribution. The functional association between regulatory regions and genes is a different category of features. Each non-coding element is associated with a set of genes depending on the value of the β factor, which determines how far the association between non-coding and coding regions will be considered (β=2 is illustrated). For a given inter-marker region, the predictor will test if the region overlaps such associations. In total, 29 features are considered.

The other type of features considered describes the relationship between non-coding conserved regions (considered in this study as enriched for *cis*-regulatory elements) and their putative target coding regions. This relationship is described as a function of a parameter β. With $\beta = 1$, each non-coding marker is linked to the gene with the closest transcription start site (up to a maximal distance of 1 Mb). For $\beta > 1$, each non-coding marker is linked to its closest gene and to all other genes located within at most β times the distance to the closest gene. Consequently, the higher the β, the more genes are linked to a single non-coding region. The 'association' feature of a given inter-marker region is then defined by the number of such associations that would be broken by a break in that region. These features have been chosen to test the hypothesis that long-range regulation may be a factor in maintaining the integrity of the genome. Under that hypothesis, breakpoints would be expected to occur where no (or few) regulator-target gene connections are broken.

2.4.3 Removing inter-marker distance bias

Unsurprisingly, the length of an inter-marker region is strongly correlated with its likelihood to contain a breakpoint (logistic regression analysis, p-value = 4.5e-6). This confounding factor needs to be factored out before more interesting predictive features can be teased out. To this end, we fitted the breakpoint/no-breakpoint binary data using a linear regression based on inter-marker fragment length (Breakpoint(r) $\sim$ $a\cdot$ Length(r) + b) and obtained the residuals of the regression (Residual(r) = (Breakpoint(r) - b)/a). Large inter-marker regions with no breakpoint produce large negative residuals, while small inter-marker regions

with a breakpoint produce large positive residuals. It is these residuals, which should be considered as fragment labels normalized for fragment lengths, that are used as target values for the predictors that follow.

2.4.4 Breakpoint predictors

We first tested the predictive value of each individual feature. Then, the best combination of features was selected with a forward feature selection procedure. Each feature was first tested independently to examine its ability to predict breakpoints, measured by the t-value of the linear regression of the length-normalized breakpoint data against that feature. A graph showing both the effect for local density features and relationship between coding and non-coding features (for different β values) is presented in Figure 2–3. A large negative t-value represents a negative correlation of the feature with the presence of breakpoints, whereas a large positive t-value indicates that the presence of this feature is favorable to breakpoints. We observe that a high local density of coding elements (both metatherian and amniote) is associated to an increased likelihood of breakpoints, corroborating our previous observation that breakpoints occur more often than expected between coding markers. This is in accordance with observations showing that the synteny of conserved non-coding elements within gene deserts is usually well conserved [Ovcharenko *et al.*, 2005]. Interestingly the density of more ancient genes (amniote) is more strongly associated to breakpoints than that of more recent ones. Indeed, as we will see bellow, most housekeeping genes are shared among amniotes and they are also associated to such breaks. In addition, we note that the value of the locality parameter α has little impact

on the fit, suggesting that the best predictive features would be a more complex function of the distance. On the other hand, non-coding markers (or putative regulatory regions) are negative predictors for breakpoints. Interestingly, although non-coding amniote and metatherian densities are strong negative breakpoint predictors when α is small, only the non-coding amniote density remain predictive for large values of α, indicating that breakpoints in the immediate proximity of such ancient regulatory elements are quite rare, but that breakpoints near more recent non-coding are less deleterious. Features modeling the association between coding and non-coding markers have a negative t-value, which means that breakpoints are less likely to become fixed in the population if it breaks such association. Finally, the best t-value obtained is for $\beta = 1.5$, indicating that the regulator/target gene relation often is not limited to a non-coding region and its (single) closest gene.

2.4.5 Predictor training and cross-validation

A multiple linear regression breakpoint predictor was built using a forward feature selection procedure, whereby we iteratively add to the predictor the feature that yields the largest accuracy improvement, until no further addition is beneficial. To test the performance of each intermediate and final predictor, we performed a four-fold cross-validation. Instead of measuring the accuracy of our predictors in terms of the fraction of inter-marker region correctly predicted to have a breakpoint, the sensitivity and specificity of each predictor was assessed in terms of the total fraction of the genome predicted to be breakpoint-sensitive. Table 2–2 reports the result of the feature selection procedure.

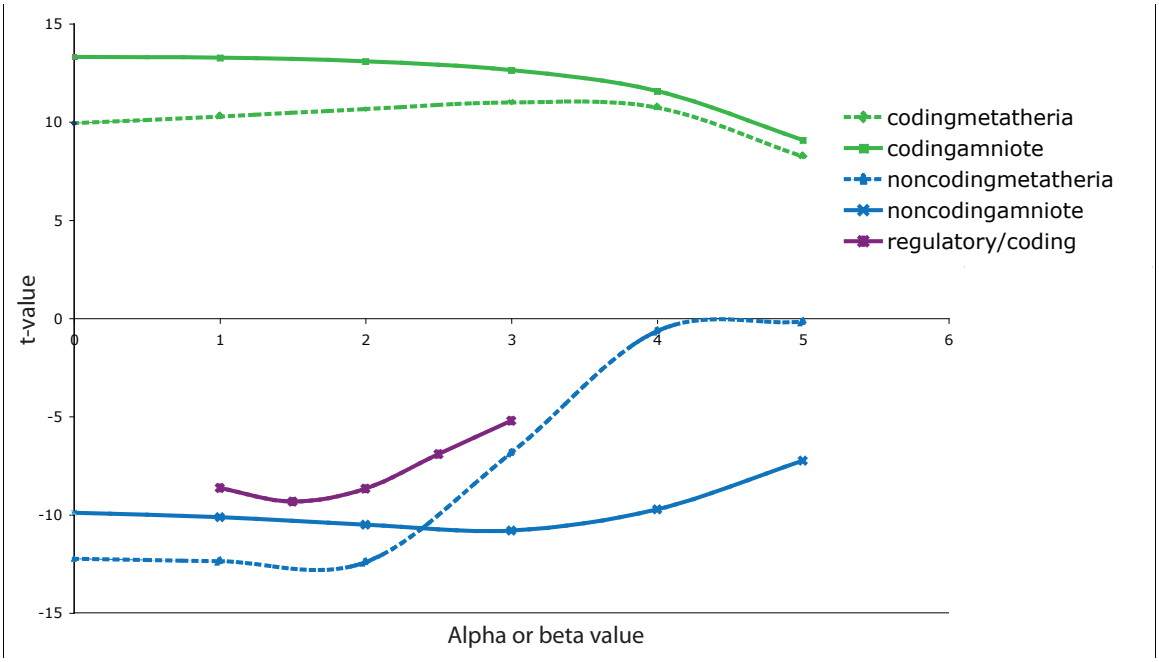

Figure 2–3: **Effect of single feature on the prediction.** t-value obtained for the linear regression of length-normalized breakpoint data for the different types of local density features (coding amniote, noncoding amniote, coding metatheria, noncoding metatheria), for different values of α. On the same graph are represented the t-values of the regression against the presence of association between conserved non-coding elements and genes, for different values of β.

Table 2–2: **Effect of each selected feature on the prediction.** Features are listed in the order in which they were selected by the forward feature selection procedure. The coefficient estimate, standard error, t-value, and p-value reported are those obtained for the linear regression with all 9 features. The specificity reported is that of the predictor built using the features starting from the 1st row down to the current row. The specificity is calculated for a sensitivity of 0.75. For example a specificity of 0.645 means that 75% of breakpoints are comprised within 35.5% of the total length of inter-marker regions used for the analysis.

Feature	Estimate	Std. Error	t value	p-value	specificity
CodingAmniote, $\alpha = 0$	0.0030645	0.0005186	5.910	3.47e-09	0.473
AssociationBreaks, $\beta = 2$	-0.0150509	0.0022701	-6.630	3.42e-11	0.572
CodingMetatheria, $\alpha = 2$	1.3915142	0.2395547	5.809	6.36e-09	0.611
NonCodingMetatheria, $\alpha = 3$	-0.0338532	0.1649918	-0.205	0.83743	0.625
CodingMetatheria, $\alpha = 0$	-0.0079558	0.0014829	-5.365	8.15e-08	0.629
NonCodingMetatheria, $\alpha = 2$	-0.0154798	0.0132948	-1.164	0.24429	0.630
CodingAmniote, $\alpha = 3$	-4.6993169	0.9579974	-4.905	9.38e-07	0.634
AssociationBreaks, $\beta = 3$	0.0069784	0.0024624	2.834	0.00460	0.645
NonCodingMetatheria, $\alpha = 4$	0.0302276	0.1104565	0.274	0.78435	0.645

Choosing the appropriate prediction score threshold, the predictor identifies 35.5% of the genome as breakpoint-prone, and these regions indeed contain more than 75% of the actual breakpoints, more than twice the expected accuracy of a random predictor. After assessment of its performance, the predictor was trained on the whole data set (using opossum as an approximation to the ancestral context) and applied to the prediction of breakpoints in the human context. Surprisingly, this predictor outperforms the original one at predicting past human breakpoints succeeding at capturing 75% of breakpoints in break prone regions covering only 27% of inter-marker regions (see Figure 2–4), indicating that either the opossum genome is not a very good approximation to the ancestral context, or that the derived state matters as much as the ancestral state to predict breakpoint fixation.

Figure 2–5 shows the breakpoint susceptibility profile of human chromosome 2, using a 500 kb sliding window. This score is strongly positively correlated with the number of coding regions and negatively correlated with the number of non-coding conserved regions. However, some regions, such as the Hox D cluster on chromosome 2, a gene rich region but known to contain several evolutionary conserved non-coding regions, is also predicted with a low score.

2.4.6 A limited fraction of the genome can tolerate breakpoints.

The predictor was applied to the complete human genome to identify regions that are more likely to tolerate breakpoints. We divided the genome into two sets of regions: *susceptible* regions are those predisposed to rearrangement (regions with score above 0.0065, covering 30% of the non-genic euchromatic genome),

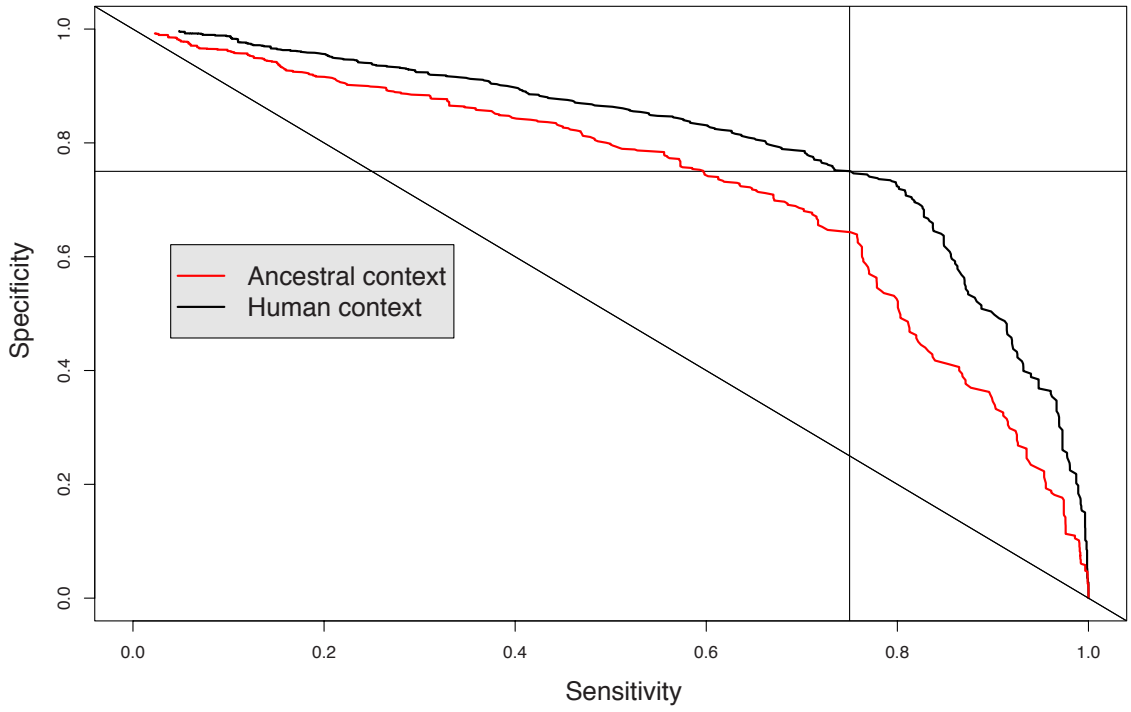

Figure 2–4: **Specificity/sensitivity curve for breakpoint predciction.** Specificity/sensitivity curves are presented for both predictions (ancestral context or the human (derived) context).

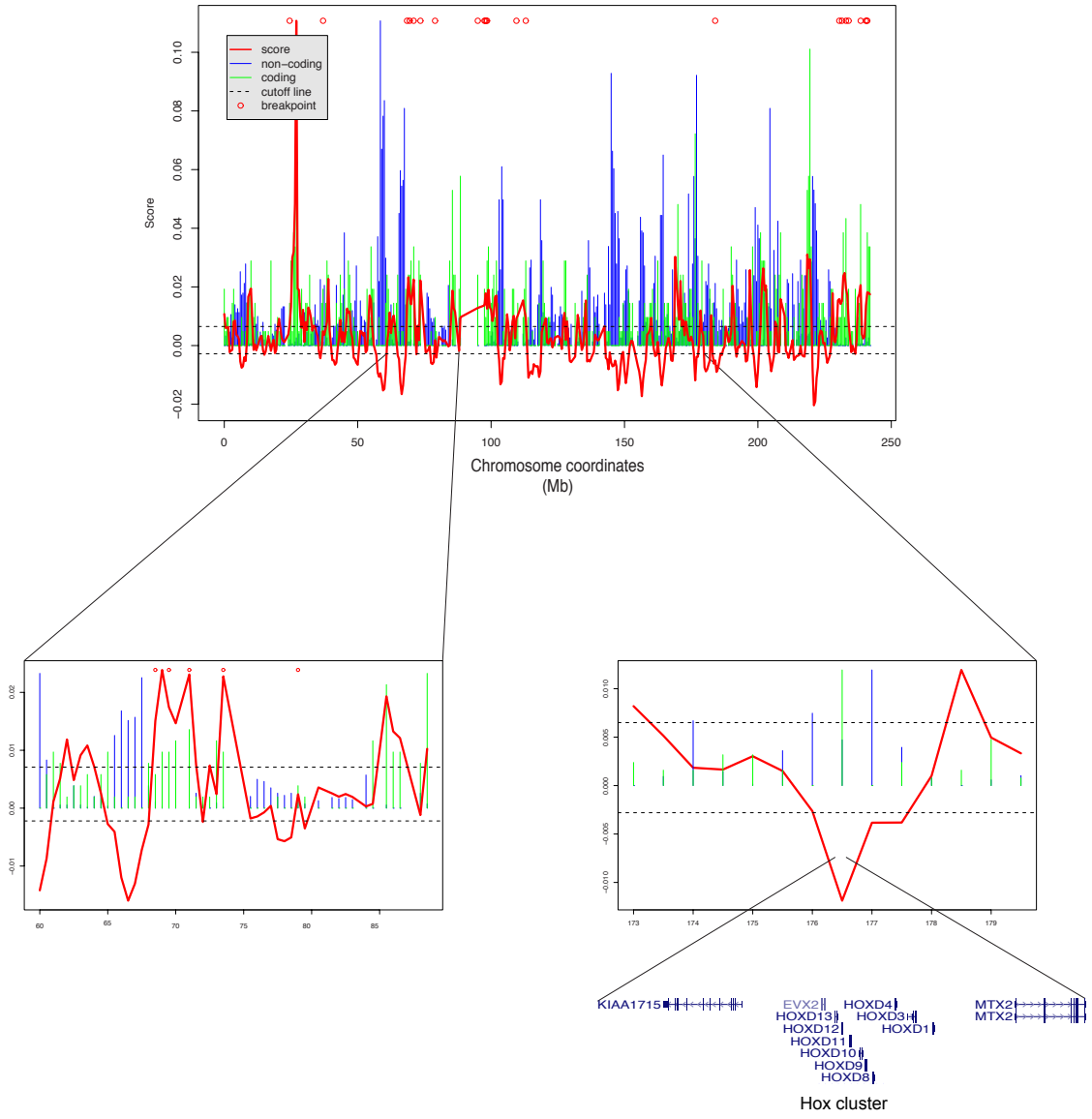

Figure 2–5: **Prediction scores on chromosome 2.** Normalized number of non-coding and coding elements are respectively represented with blue and green bars. Scores are represented by a red continuous line. The bottom right part of the figure show the region corresponding to chromosome 2 Hox cluster. The bottom left part shows a region of chromosome 2 with a high density of breakpoints.

and *refractory* regions, referring to those that are resistant to rearrangements (regions with score below -0.0028, covering 30% of the non-genic euchromatic genome). About 73% of human breakpoints are comprised in susceptible regions (a 2.3-fold enrichment), while only 7% are found within refractory regions (a 4.3-fold depletion). Most of the breakpoints are then contained in a limited fraction of the genome. This clearly shows that breakpoint fixation is not happening randomly and uniformly across the genome and that regions that are more likely to be broken can be predicted. Moreover if we consider that breakpoints almost never occur within genic regions or within conserved regions, we obtain that more than 73% of the human breakpoints are located in about 20% of the genomic regions considered. This observation complements the theory of evolutionary hotspot described by Peng *et al.* [Peng *et al.*, 2006]. We then used this classification to uncover additional properties of each type of regions.

2.4.7 Susceptible and refractory regions have different characteristics

Susceptible and refractory regions differ in a number of aspects.

1. **Refractory regions are strongly enriched for non-coding markers, and susceptible regions for coding markers**

   The ratio of coding to non-coding markers is significantly higher in susceptible regions than in refractory regions (p-value $< 2^{-16}$, Fisher test, see Table 2–3). This result meets observations made by Murphy *et al.*, showing that there is a significant increase of gene density in breakpoint regions [Murphy *et al.*, 2005].

2. **Refractory regions are enriched for Trans/Dev genes**

Table 2–3: **Properties of refractory and susceptible regions.**

	Refractory regions	Susceptible regions	Refr./Susc. Ratio
Coding markers	1484	7423	0.20
Noncoding markers	41947	5576	7.5
Specific genes	374	1636	0.23
Ubiquitous genes	194	1485	0.13
Gene deserts	142	5	35.5

A Gene Ontology analysis (performed on the "biological process" classification with the Babelomics platform [Al-Shahrour *et al.*, 2008]) reveals that refractory regions are strongly enriched for genes involved in development, such as anatomical structure development (p-value 1 x 10^{-10}), multicellular organismal development (p-value 9 x 10^{-12}) and regulation of biological process (p-value 2 x 10^{-6}) (see Figure 2–6). This confirms the observation made that developmental genes are enriched in syntenic regions [Engström *et al.*, 2007]. Interestingly, susceptible regions are enriched for genes involved in immune response. These genes must be extremely adaptive and genes such as immunoglobulin are under intense gene diversification processes such as gene conversion, somatic hypermutation and class switch recombination [Maizels, 2005]. It is then not surprising to predict higher rearrangement rates in regions involved in immunity which are under strong positive selection pressure. However, this may also be an artifact caused by the intense duplication history of some of these genes, which makes them more susceptible to misalignment.

3. **Refractory regions are enriched for tissue specific genes.**

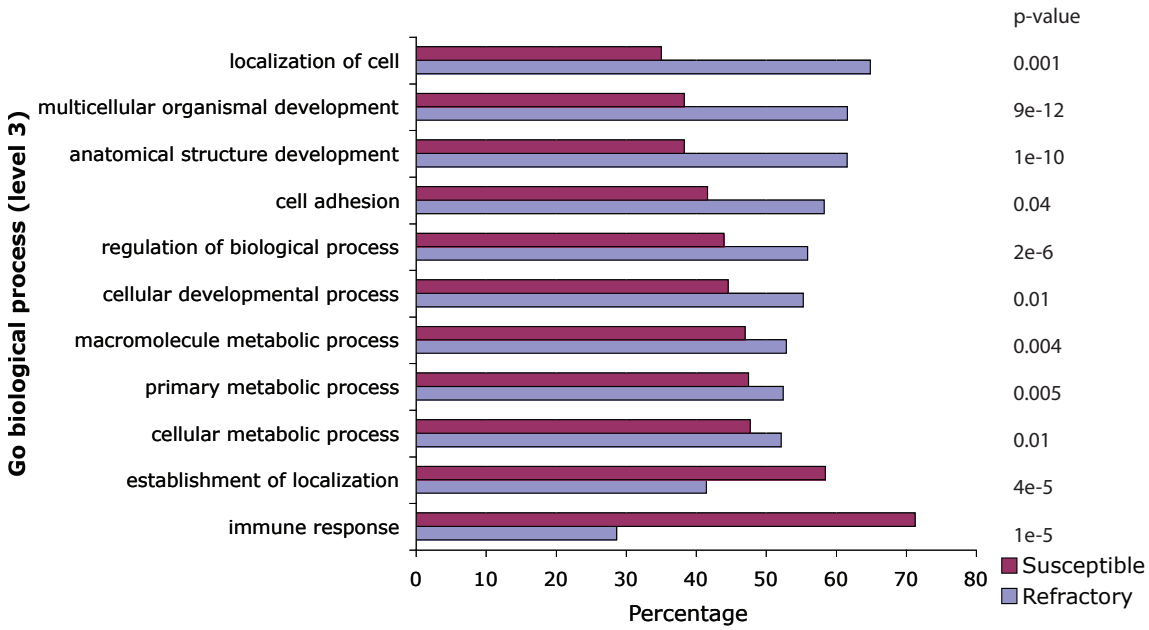

Figure 2–6: **GO categories enrichment and depletion in susceptible and refractory regions.** Over- and under-represented GO categories (biological process, level 3 only) for genes localized within susceptible and refractory regions. The adjusted p-values were obtained with a two-sided Fisher test using the Fatigo function from the Babelomics platform.

We used the GNF Expression Atlas 2 [Su *et al.*, 2004] to classify the human genes based on their expression in 79 human tissues and cell types. The dataset contains expression measurements for 14 614 distinct Ensembl genes. Each gene was classified according to the number of tissues in which it is expressed [Lifanov *et al.*, 2003]. A gene is considered expressed if the detected expression level is above a certain threshold. Using this classification method, two gene sets were created. The set of 'specific' genes consists of all genes expressed in at most 5 tissues and contains 3 235 genes

(see Methods). The set of 'ubiquitous' genes contains 2 520 genes expressed in more than 70 tissues. The remaining genes expressed in 6 to 69 tissues where not used for this analysis. The ratio between the number of specific genes and ubiquitous genes is clearly imbalanced between refractory and susceptible regions. Refractory regions are clearly enriched for specific genes compared to susceptible regions (two-sided Fisher-test, p-value 3 $\times 10^{-9}$, see Table 2–3).

4. **Most gene deserts lie in refractory regions.**

   About 25% of human genome is composed of gene deserts, which are defined as long inter-genic regions [Venter *et al.*, 2001]. In this work, we define a gene desert as a genomic region of more than 1Mb without protein coding genes. The human genome contains 270 gene deserts, of which 142 fall within refractory regions (based on their average score) but only 5 within susceptible regions (the 123 others are located in regions that are neither predicted as refractory nor susceptible). This result agrees with our previous observation that most breakpoints avoid non-coding conserved regions and is consistent with previous studies stating that most gene deserts are not broken by evolutionary breakpoints [Ovcharenko *et al.*, 2005].

   The predicted scores in gene deserts follow a bimodal distribution, as shown Figure 2–7. From this distribution, we can distinguish two types of gene deserts: (i) those whose score is under the refractory threshold, where evolutionary breakpoints are not likely to happen and (ii) those over this threshold. Interestingly, this dichotomy of gene deserts for susceptibility

to breakpoints is somewhat similar to observation made by Ovcharenko *et al.* [Ovcharenko *et al.*, 2005], who noted that gene deserts can be separated into two kinds: 'stable' and 'variable'. Stable and variable gene deserts are described with different properties: genes flanking stable gene deserts are enriched for transcriptional and developmental functions and are resistant to rearrangements. This is consistent with our observations on susceptible and refractory regions.

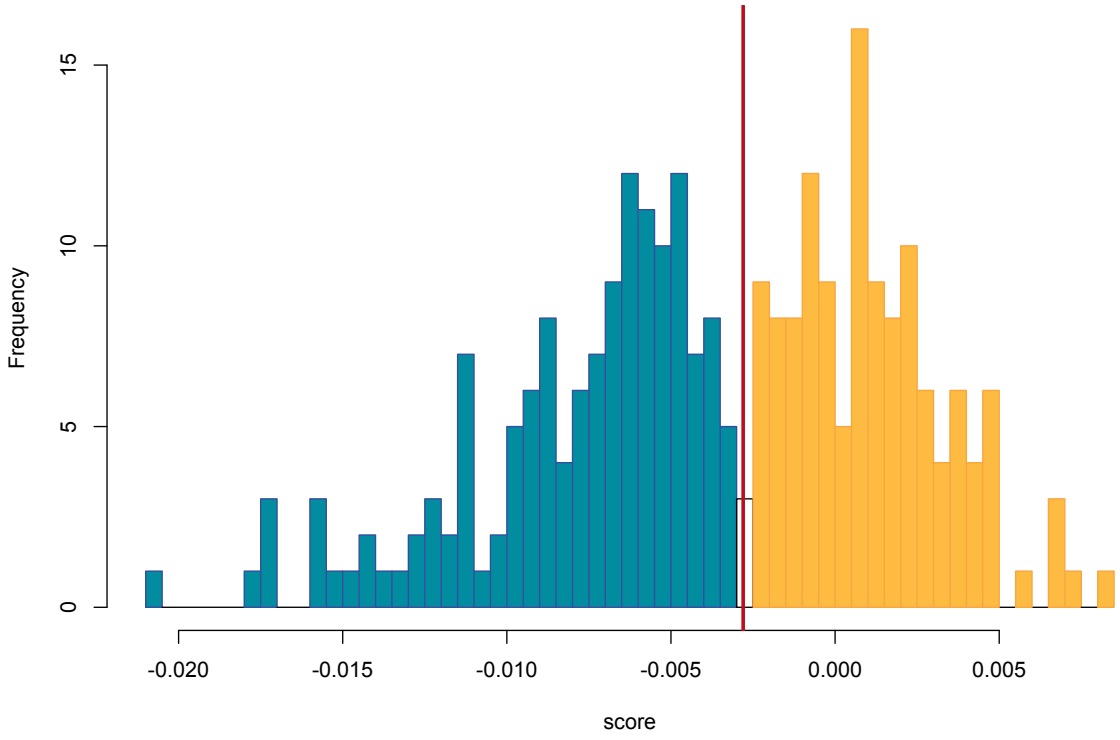

Figure 2–7: **Distribution of the average prediction score of gene deserts.** The red line corresponds to the threshold below which a region is considered refractory. The distribution is clearly bimodal, as highlighted by the coloring scheme.

5. **Susceptible regions are enriched for copy number variations**

   Copy number variations (CNVs) are regions (1Kb to 1Mb) of the human genome whose copy number is polymorphic in the population. From the database of genomic variants [Iafrate *et al.*, 2004], we retrieved the base pair coverage for copy number variations in susceptible and refractory regions. As the database contains various kinds of variation such as inversions, we selected only variations labelled as copyNumber. CNVs are significantly enriched in susceptible regions, compared to refractory region. 25.1% of base pairs in susceptible regions are covered by CNVs, whereas this is the case for only 19.6% of those in refractory regions (see Table 2–4). Regions with high coverage of CNVs are indeed regions of the genome where variations in size are potentially less detrimental. It is not surprising then to find an enrichment for CNVs in regions predicted as susceptible to breakpoints. This CNVs analysis is another independent confirmation of the validity of our predictor.

Table 2–4: **Percentage of susceptible, neutral and refractory regions covered by rare, common fragile sites and CNVs.** The statistical significance of the enrichment between refractory and susceptible regions for rare and common fragile sites as well as CNVs regions was assessed with a permutation test. The resulting p-values are $< 10^{-4}$

	% covered by rare fragile sites	% covered by common fragile sites	% covered by CNVs
Susceptible regions	8.8	24.7	25.1
Neutral regions	4.5	22.3	22.3
Refractory regions	5.5	20.7	19.6

6. **Susceptible regions are enriched for rare fragile sites**

Fragile sites are regions of the genome which appear as gaps or breaks on metaphase chromosome when exposed to inhibitors of DNA synthesis. Those regions are considered as 'unstable' part of the chromosome [Durkin & Glover, 2007]. Fragile sites are further categorized depending of their frequency: rare fragile sites are present in a small proportion of individuals whereas common fragile sites are present in all individuals and are considered part of the chromosome structure [Durkin & Glover, 2007, Ruiz-Herrera *et al.*, 2006]. We used 119 fragile sites (88 defined as common and 31 as rare) reported by Schwartz *et al.* [Schwartz *et al.*, 2006].

Rare fragile sites are clearly enriched in susceptible regions compared to neutral and refractory regions, see Table 2–4. Those data agree with observations already made by Ruiz-Herrera *et al.* [Ruiz-Herrera *et al.*, 2006] who showed a weak correlation between common fragile sites and evolutionary breakpoints and a more significant correlation between evolutionary breakpoints and rare fragile sites. We should point out the difference of resolution between the cytogenetic bands representing fragile sites (which are on average 7Mb long) with our estimate of susceptible and refractory regions, which is more refined.

2.5 Discussion

2.5.1 Breakpoints are bound to specific regions

In this study, we developed a predictor to define regions of the human genome that are likely to tolerate rearrangements. Using this predictor, we defined two

classes of regions. Susceptible regions correspond to 30% of the intergenic genome and contain 73% of the breakpoints. Refractory regions correspond to 30% of the intergenic genome but contain only 7% of the breakpoints. Most breakpoints are then contained in a small portion of the genome. Considering that coding loci are also extremely refractory to breakpoints, only 20% of the human regions considered for the analysis are prone to rearrangements. This model - that breakpoints are concentrated in a small, identifiable fraction of the genome - complements the 'fragile breakage' model proposed by Pevzner and Tesler [Pevzner & Tesler, 2003], which was developed as an alternative model to the random breakage theory introduced by Nadeau and Taylor [Nadeau & Taylor, 1984].

2.5.2 Long-range regulation imposes functional constraints on the genomic structure

Regulatory regions and genes can be functionally associated over long stretches of DNA. Some regulatory regions have indeed been located as far away as 1Mb away from their target genes (Shh long-range enhancer, for example [Lettice *et al.*, 2003]). Vavouri *et al.* also showed using duplicated conserved non-coding elements and paralogous genes that about half of non-coding elements are $>$ 250kb away from their target gene [Vavouri *et al.*, 2006]. This long-range interaction associated with the complex relationship between regulatory regions and target genes (a gene can be targeted by many regulators and a regulator can target many different genes) establishes an important pressure to keep those regulators and target genes together. Intuitively, the cost of breaking the physical relationship between long-range regulatory regions and their target genes may be so high that rearrangements are rarely fixed where such relationship is

predominant (see also [Becker & Lenhard, 2007]). Our analysis of susceptible and refractory regions sheds light on the validity of this hypothesis.

2.5.3 Susceptible and refractory regions are functionally different

Refractory and susceptible regions have many distinguishing features: (i) refractory regions are significantly enriched for putative regulatory regions and gene deserts; (ii) the ratio of housekeeping genes to cell type/tissue specific genes is higher in susceptible regions than in refractory regions; (iii) refractory regions are clearly enriched for genes involved in transcriptional regulation and developmental processes (trans/dev genes). Those distinct features show a functional dichotomy between susceptible and refractory regions, between regions that are involved in complex processes (e.g. transcriptional regulation of developmental genes) and regions enriched for housekeeping genes and depleted for non-coding conserved regions. This dichotomy is in our point of view a strong argument supporting the hypothesis that long-range regulation imposes constraints on the genomics structure. This confirms previous observations where synteny blocks overlap regulatory domains [Engström *et al.*, 2007]. For example, transcription factor genes - enriched in refractory regions - are under complex regulation and can be expressed at different levels, at different times and in different tissues [Wray, 2007]. We also showed that copy number variations - an independent dataset - are enriched in susceptible regions in comparison to refractory region. If it does not bring any information on the cause of the instability, it however can be interpreted as the result of a reduced constraints on genome structure which could be due to decreased regulation complexity.

2.5.4 Reduced regulation complexity: cause or consequence of breakpoint susceptibility?

It has been shown that there is a significant overlap between evolutionary breakpoints and fragile site locations and that, even if no mechanistic role could be demonstrated, some fragile regions of the genome may be more likely to experience reorganization [Ruiz-Herrera *et al.*, 2006]. One may then wonder whether the relative regulatory simplicity observed in susceptible regions may actually be a consequence (rather than a cause) of the presence of fragile regions nearby. But although fragile regions are correlated with regions susceptible to breakpoints as shown by our data, they only represent a small fraction of susceptible regions, suggesting other mechanisms explaining those different levels of plasticity on the genome. If fragile regions may contribute to the fixation of breakpoints, it seems that the main mechanism preventing breakpoints is the crucial role that long-range regulation has on the fitness of the individuals.

2.5.5 Limitation of the model and further developments

The predictor was trained using the opossum genome as an approximation for the ancestral eutherian genome. Surprisingly the predictor performs better when using the current human genome (i.e. the derived genome), rather than the approximated ancestral genome, for predicting human breakpoints (see Figure 2–4). This outcome may be explained by the discrepancy in the quality of assembly and annotation between human and opossum. Nonetheless, we believe that using opossum as an approximation of the ancestral state is justified as the alternative, using a computationally predicted ancestral genome, may lead to a worse approximation because of reconstruction errors. Another observation to make from this

discrepancy is that the derived state may be as suitable as the ancestral state to predict breakpoints. Indeed, the effect of a genome rearrangement is a combination of both the regulatory associations it disrupts (observable in the ancestral genome) and the new associations it creates (observable in the derived genome).

2.6 Conclusions

We show in this study that the reason why some regions of the genome are not prone to rearrangement is that some of the genes they contain are under the influence of long-range regulators and the physical relationship between these elements cannot be broken without being detrimental to the fitness of the individual. Genes with simpler regulation, such as housekeeping genes, may be less affected by breakpoints in their surroundings. The consequence is that regions where rearrangements can be fixed and are not too detrimental correspond to regions that are enriched for genes with less complex regulation. In the light of these data, we confirm that the random breakage model is not the most appropriate and that only a limited fraction of the genome is susceptible to evolutionary rearrangements. The mapping of these regions, produced by our predictor, will be of importance for future genome evolution and function studies.

2.7 Methods

2.7.1 Marker identification

In order to undertake our analysis on the synteny of human putative regulatory regions and coding regions, we defined both datasets from publicly available data. As it is widely accepted that non-coding regions under selective pressure are enriched for regulatory regions, we selected the set of non-coding conserved regions

from the human 28-way [Miller *et al.*, 2007] alignment identified by PhastCons [Siepel *et al.*, 2005] and available on the UCSC genome browser [Kent *et al.*, 2002]. Only regions longer than 50bp and with a score over 400 (third quartile from the complete distribution of Phastcons elements) were considered. These regions were filtered out for ESTs, coding regions (exons), blastp hits and repeats using Ensembl annotations. Coding regions are defined from the set of human coding exons from the Ensembl version 49 [Hubbard *et al.*, 2007]. When two exons overlap (which occurs in the case of splice variants), only the longest exon was considered. In the case of two overlapping genes (e.g, intronic gene), only the longest gene was taken into account. Through this process, we selected 216 300 exons (coding markers) and 112 964 non-coding conserved regions (non-coding markers) to undertake the analysis.

2.7.2 Ortholog mapping

Based on whole genome 'liftover chains', pairwise alignments (human/opossum and human/chicken) were retrieved from the UCSC genome browser [Kent *et al.*, 2002]. Liftover chains were extracted from UCSC nets generated from blastZ alignments. Nets are a hierarchical collection of ordered aligned blocks and the mapping provided by this alignments is then unlikely to be spurious. Human (NCBI build 36.1) is used as the reference genome, and human conserved regions (coding and non-coding) were mapped using liftOver (forward mapping) to the chicken v2.1 draft assembly (WUSTL) and the opossum draft assembly (The Broad Institute, January 2006) (liftOver parameters: minMatch=0.8 for opossum, minMatch=0.7 for

chicken). In order to only consider best reciprocal hits, the forward mapping results are mapped back to human (reverse mapping) using liftover (minMatch=0.75 for opossum, minMatch=0.65 for chicken). Markers lying on unknown chromosomes of the human genome or mapping to unknown chromosomes on one of the target genomes were discarded. Each marker (coding or non-coding) was then classified using its level of conservation. Markers conserved only between human and opossum were classified as metatherian. Those conserved between human and both chicken and opossum are classified as amniote markers (we ignored markers that were conserved only between human and chicken).

2.7.3 Synteny

A breakpoint between the human genome and the opossum genome (with respect to the chicken genome) was defined as a pair of amniote markers (coding or non-coding) that are adjacent in human (disregarding possibly intervening metatherian markers) but not in opossum (with respect to chicken). The only exception is that if the two markers are more than 1Mb apart, the inter-marker region is disregarded, as in that case the resolution would be too low. This removes from consideration cases such as centromeres. As the exact breakpoint position cannot be determined, its localization is defined as the equidistant position between the two markers.

In order to undertake analyses at the gene level, exons were assembled into genes using the Ensembl annotation, and synteny breakages are ported at the gene level (placing the breakpoint on the left side of the gene. Marker classification (amniote and metatheria) was also ported from exons to genes. If at least 30% of

the exons were labelled as amniote, the amniote annotation is ported to the gene. If at least 30% of the exons were labelled as metatheria (and less than 30% are labelled as amniote), the annotation metatheria was ported to the gene. Finally, those breaks retrieved on human were mapped to the opossum genome where the predictor training was undertaken.

In order to evaluate the significance of the enrichment of observed breakpoints depending of the flanking markers, we calculated the number of expected breakpoints based on the total size of inter-marker regions of each type using a chi-square test.

2.7.4 Breakpoint prediction

Inter-marker regions were divided into two classes; syntenic regions and breakpoint regions. To train the predictor, the following information was used: local density of functional elements and association between putative regulatory regions and genes. A score summarizing the local density of elements within 1Mb of the center of each inter-marker region was considered. The following elements were considered: the status of markers (coding or non-coding), their classification (metatheria or amniote), and their weighted distance. For a 2Mb window W centered at genomic position p, feature scores were calculated as shown equation 2.1 where $X \in \{\text{CodMet, NoncodMet, CodAmn, NoncodAmn}\}$ and α ranging from 0 to 5 (with increments of 1).

$$F_X(p,\alpha) = \sum_{\text{marker } m \text{ of type } X \text{ in } W} \frac{1}{\log(|pos(m) - p|)^\alpha} \tag{2.1}$$

Another feature considered is the connectivity between non-coding putative regulatory regions and genes. All non-coding conserved regions associated to the gene with the closest transcription start site. In addition, a non-coding region was associated to a gene if the distance between them was at most β times more than the distance to the closest gene, for β ranging from 1 to 3 with 0.5 increments. For a given inter-marker region centered at position p, we then calculated $F_{assoc}(\beta)$, the number of associations that cross position p (i.e. associations that would be destroyed by a breakpoint).

A logistic regression was first applied on the data with inter-maker distance as a unique predictive feature. Residuals obtained from the regression were then used to train a multiple linear regression predictor. The use of the residuals allowed the capture of information that is not related to this inter-marker distance.

The 29 features are composed of four types of markers (coding metatheria, non-coding metatheria, coding amniote and non coding amniote) analyzed with 6 different α values and 5 different β values representing putative associations between a non-coding conserved region and a gene. The 29 features were first tested separately as single predictors and their effect on the prediction assessed with the t-value associated with the linear predictor output. Then, features were selected using a forward selection method. Each addition of a new feature was selected using the highest specificity value for a given sensitivity of 0.75. Sensitivity and specificity were calculated using the number of base pairs covered and not the number of inter-marker distance. We then undertook a four-fold cross

validation on the oppossum genome and the predictor was finally applied on the human genome where further functional analysis was undertaken. In addition, we tried to add interactions between features but this didn't bring much improvement.

2.7.5 Additional datasets

Gene deserts are defined as gene-free regions spanning more than 1Mb, based on the Ensembl gene annotation version 49. Only genes labeled as "known" were used. All regions with more than 1/3 of non-sequenced base pairs were removed from the dataset.

GNF Expression Atlas 2 [Su *et al.*, 2004] allows classifying genes depending on their expression in the 79 human organs and tissues covered by the Atlas. We considered that a gene is expressed in a given tissue if its MAS5 normalized expression level is > 400. We ported the GNF microarray probes to the Ensembl geneset using the Biomart tool [Kasprzyk *et al.*, 2004] on the Ensembl web site.

2.8 Acknowledgements

The authors wish to acknowledge support from Genome Quebec/Canada. We also wish to thank Francois Spitz from EMBL Heidelberg and Rob Sladek from McGill University for useful discussions. We also thank the reviewers for their helpful comments and suggestions.

CHAPTER 3
Mapping associations between long-range *cis*-regulatory regions and their target genes using synteny

3.1 Preface

In **Chapter 2**, we observed that long-range regulation may be a major driving force in maintaining genome integrity. The finding resulting from that work motivated this subsequent investigation. The approach presented in this chapter aims at predicting functional interactions between *cis*-regulatory regions and putative target genes, or in other words what *cis*-regulatory regions regulate which gene. The rationale behind the approach is based on the hypothesis that rearrangements preferentially occur in regions that would not affect the fitness. The cost of breaking functional interactions between genes and regulatory regions would be too high and consequently not fixed in evolution. Therefore, we developed a method that studies rearrangements between regulatory regions and genes to assess the likelihood of a pair composed of a regulatory region and a gene to functionally interact. This Chapter corresponds to a manuscript in preparation that will be submitted as follows:

- **Emmanuel Mongin**, Ken Dewar, Mathieu Blanchette. Mapping association between long-range *cis*-regulatory regions and their target genes using synteny.

3.2 Abstract

Long-range *cis*-regulatory regions are involved in the control of transcription initiation (as repressors or enhancers). Their main characteristics are: i/ that they can be located as far as 1 Mb from the transcription start site of the target gene, ii/ they can regulate more than one gene, iii/ they are usually orientation independent. Therefore, the identification and proper characterization of functional interactions between long-range *cis*-regulatory regions and target genes remains problematic.

We present a novel method to predict such interactions based on the analysis of rearrangements between human and 16 vertebrate genomes. Our method is based on the assumption that genome rearrangements are likely to be deleterious if they disrupt the functional interaction between a *cis*-regulatory region and a candidate target gene. Therefore, conservation of synteny through evolution may be an indication that the pair members may functionally interact.

In this study, we propose an Expectation-Maximization algorithm that classifies putative enhancer/gene associations as functional or non-functional based on their evolutionary history. We use our algorithm to classify a set of 1,406,084 putative associations from the human genome, based on the comparison to 16 other vertebrate genomes.

This genome-wide map of interactions, has many potential applications among which are the selection of candidate regions prior to *in vivo* experimental characterizations; a better characterization of regulatory regions involved in

position effect diseases; or to shed light on the mechanisms and importance of long-range regulation.

3.3 Introduction

The regulation of transcription is controlled by various distinct regions. Transcription initiation is, in vertebrates, controlled by distinct genomic regions that act as binding platform for transcription factors. In vertebrates, accurate regulation is crucial to many biological processes such as development, tissue specificity, or response to external stimuli. Distinct *cis*-regulatory regions, with their own specificities, take part in complex cross-talking processes that result in proper gene regulation. However, alteration of those regions or disruption of the physical link between *cis*-regulatory regions and target genes can have dramatic phenotypic effects often leading to diseases ([Leipoldt *et al.*, 2007], [Trembath *et al.*, 2004]). Among those classes of regulatory regions, we distinguish: i/ the core promoter, bound by the transcription initiation complex, responsible for the initiation of transcription; ii/ the proximal promoter, usually defined as the region up to 1.5 kb upstream from the transcription start site (TSS). (This proximal promoter region is bound by co-activators that directly interact with the core promoter to facilitate the recruitment of the basal transcription machinery [Ptashne & Gann, 1997]. This results in enhanced transcription and tissue specificity [Zhao *et al.*, 2007].); iii/ long-range *cis*-regulatory regions that are located over 1.5 kb upstream or downstream from the TSS and regulate

genes at distances reaching 1Mb [Lettice *et al.*, 2003] (Shh gene). (These long-range regulators comprise various type of regions such as enhancers, silencers or insulators.)

Long-range regulatory regions may regulate many genes [Spitz *et al.*, 2003]; can act in an orientation independent manner [Atchison, 1988]; and more importantly can regulate target genes as far as 1 Mb. Consequently, for a given *cis*-regulatory region, predicting its putative gene target(s) is a difficult task. Such predictions would meet with various interests: i/ to determine what transcription factor regulate what genes and establish regulatory networks; ii/ to better understand the mechanisms of long-range regulation; iii/ to make educated choice prior to *in vivo* testing of *cis*-regulatory regions; iv/ to find the most likely regulatory regions that may be involved in position effect related diseases (chromosomal breakpoints linked to diseases but not occurring within a gene); v/ to help associate *cis*-regulatory genetic variation to expression variation (for example in studies such as [Ge *et al.*, 2009]).

Following the publications of the human genome sequence [Lander *et al.*, 2001, Venter *et al.*, 2001], we have witnessed an increasing number of vertebrate genome sequencing projects reaching completion for genomes ranging from teleosts to mammals [Aparicio *et al.*, 2002, Kasahara *et al.*, 2007, Waterston *et al.*, 2002, Gibbs *et al.*, 2004]. This, in combination with fast and accurate genomic DNA alignment programs [Schwartz *et al.*, 2003, Brudno *et al.*, 2003a, Blanchette *et al.*, 2004, Paten *et al.*, 2008] have empowered the field of comparative genomics. Of specific interest in our context are methods allowing the identification of about 100,000

non-coding evolutionarily conserved regions, most of which are theorized to be regulatory regions, yet most of which are located very far from any annotated transcript. Although there is now ample evidence linking these regions to regulatory functions [Nobrega *et al.*, 2003, de la Calle-Mustienes *et al.*, 2005, Pennacchio *et al.*, 2006], computational approaches have rarely attempted to predict the gene targets of these regulatory regions.

Evolutionary rearrangements and long-range regulation. Studies of genome evolution and genome rearrangements have also greatly benefited from the increase in genomic data. Of particular interest is the fact that evolutionary rearrangements (those rearrangements that become fixed in a population during evolution) have been shown to occur in specific regions termed "fragile regions" [Pevzner & Tesler, 2003, Peng *et al.*, 2006] and not randomly as previously modeled [Nadeau & Taylor, 1984]. The consequences of this discovery, especially in the light of long-range regulation, are valuable. The likelihood of a genome rearrangement becoming fixed in a population is strongly dependent on the fitness of the mutated individuals. In the context of long-range regulation, a rearrangement disrupting the physical link between a regulatory regions and its target gene (i.e. involving a breakpoint between the two loci) will usually be deleterious to some extent, and would rarely be fixed in evolution. Therefore evolutionary rearrangements are thought to largely involve breakpoints located in regions of the genome where they will not disrupt long-range regulation [Mongin *et al.*, 2009, Larkin *et al.*, 2009].

A new method to assess functional interaction between *cis*-regulatory region and putative target genes. In bacterial genomes, the conservation of association between groups of genes across different species has been used to predict operons (set of genes transcribed polycistronically, thus under the effect of the same regulatory region) [Huynen *et al.*, 2000, Ermolaeva *et al.*, 2001]. However, the eukaryotic regulation of transcription is more complex and understanding the functional link between specific regulatory regions and genes requires different methods. Previous studies used synteny to define the regulatory range of *cis*-regulatory regions for specific loci [Flint *et al.*, 2001] or over the whole genome [Ahituv *et al.*, 2005]. Other methods were developed to find target genes to regulatory regions but were limited to small datasets [Vavouri *et al.*, 2006, Sun *et al.*, 2008].

We propose a new computational method based on the study of the conservation of the physical association of 1,406,084 human gene/*cis*-regulatory candidate pairs in 16 other vertebrate genomes to assess the likelihood of functional interaction for each. The result is a genome-wide map of predicted functional interactions between long-range *cis*-regulatory regions and their putative target genes in the human genome by providing association scores. Contrasting with other methods, our predictions are not limited to amniote or pan-vertebrate conserved putative *cis*-regulatory regions and provides an approach that is easily scalable to new genomes.

3.4 Results

3.4.1 Orthology mapping

We assess the functional interaction between genes and putative *cis*-regulatory regions based on the conservation of their physical proximity on chromosomes in various vertebrate genomes. The gene set is composed of 25,575 human genes (EnsEMBL genes version 54 excluding pseudogenes [Hubbard *et al.*, 2009, Curwen *et al.*, 2004]), consisting of a total of 257,985 human exons. The set contains 21,404 protein coding genes, 1664 miRNAs, 1334 snRNAs, 717 snoRNAs, and 444 rRNAs.

Various functional studies have shown that non-coding regions under purifying selection are enriched for elements with regulatory properties [Nobrega *et al.*, 2003, de la Calle-Mustienes *et al.*, 2005, Pennacchio *et al.*, 2006]. Our set of putative *cis*-regulatory regions is composed of 123,905 human non-coding conserved regions (99,512 intergenic and 24,393 intronic) from the UCSC 28-way conserved regions [Miller *et al.*, 2007, Siepel *et al.*, 2005] (see Methods). In this paper, we work under the assumption that these non-coding conserved elements (NCEs) have a regulatory function, although a small fraction of them is expected to have other functions.

Genomes were selected for this analysis based two criteria. (i) Evolutionary distance from human: Highly diverged species have undergone more genome rearrangements and are thus more informative for this study. (ii) Genome coverage: Complete and accurate genome assemblies are required to assess synteny conservation; genomes sequenced at low coverage were thus excluded. The 16 species

selected (see Supplementary Table 3–5) include 8 mammals, 2 birds, one reptile, one batrachian, and 5 fish. We next mapped both types of human elements (exons and NCEs) to these genomes by taking advantage of whole genome alignments ("liftover chains" [Kent *et al.*, 2002]; see Methods). As evolutionary distance from human becomes greater, an increasing fraction of human elements fail to map to other genomes, either because they simply do not exist there or because they have diverged beyond recognition. As expected, protein coding genes exhibit a deeper overall conservation level than other types of transcribed or putative *cis*-regulatory regions (Supplementary Table 3–5). For example, 47 to 54 % of human protein coding exons map to a teleost fish whereas only 1-7% of non-coding RNAs and only 2% of non-coding conserved regions map to these species. For our analyses, the level of conservation of each gene and NCE was defined as the ancestral node corresponding to the last common ancestor (e.g. eutherian, amniote, or gnathostomate ancestor) of the set of extant species where it exists (See Supplementary Figure 3–7).

3.4.2 A map of functional interaction between regulatory elements and target genes

Our algorithm to identify functional NCE-gene associations is summarized in Figure 3–1. First, only the 1,406,084 pairs of human NCE and genes separated by at most 1 Mb are considered as potentially functional. A pair of a gene and a NCE region is labelled *associated* in a given species S: i/ if both regions have been mapped to S, ii/ they lie on the same chromosome, and iii/ they are within at most D_S bp from each other, where D_S is a species-specific distance threshold analogous to the 1Mb threshold for human but scaled based on the size of the

genome of S (see Methods). Otherwise, a pair can be either *separated* (both components exist but have been separated by a rearrangement, or only the NCE remains conserved) or *incomplete* (either the NCE or both elements could not be mapped to S).

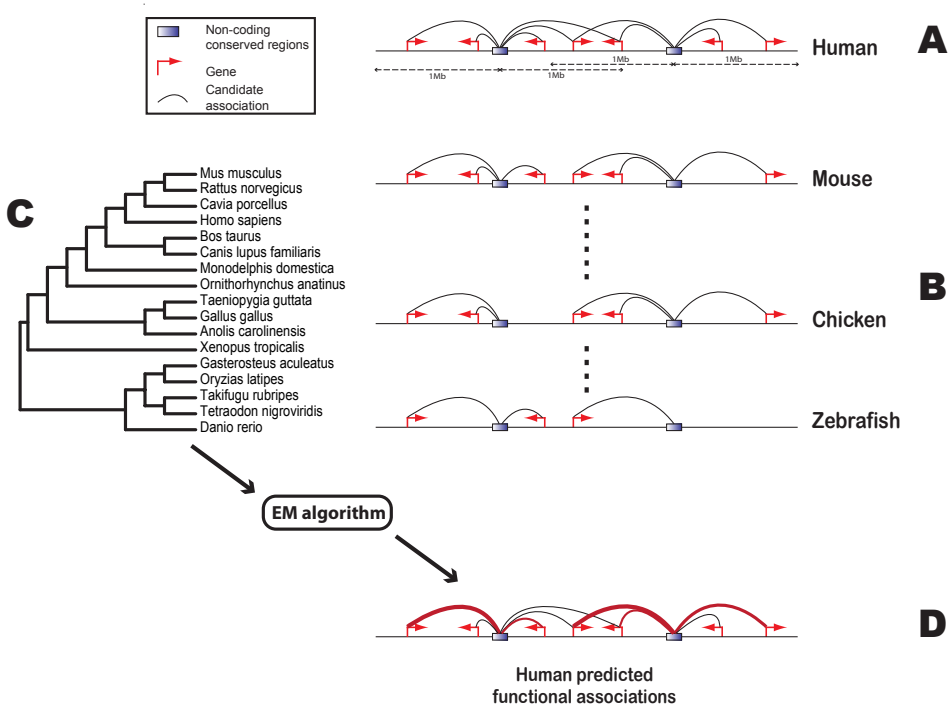

Figure 3–1: **Steps undertaken to calculate functional interaction scores.**(A) All pairs composed of a human NCE and a human gene in the same physical proximity are retrieved (candidate associations). (B) The physical association of each candidate pair is assessed in the 16 vertebrate genomes. (C) For each pair in each genome a phylogenetic tree of association is reconstructed with the Fitch algorithm. (D) Final scores are calculated from those trees with an Expectation/Maximization algorithm.

From the mapping data, the ancestral association status (associated, separated, or incomplete) of each pair is first inferred for each ancestral node of a phylogenetic tree using a variant of the Fitch algorithm ([Fitch & Margoliash, 1967]). An expectation-maximization (EM) algorithm is subsequently used to learn (in an unsupervised manner) two models: one for functionally associated pairs, and one for non-associated pairs (see Figure 3–1, sections C and D). Each model specifies the probability of maintaining or breaking association at each node of the phylogenetic tree. The functional interaction score for each pair is obtained as the log-likelihood ratio of the two models.

The distribution of functional interaction score is tri-modal (see Figure 3–2, A). The first peak (score < -10) includes 327,511 pairs for which functional interaction can be clearly ruled out based on evolutionary evidence - we call these pairs *confidently non-associated*. Pairs scoring from -10 to 49 belong to a grey zone where evolutionary evidence is inconclusive, and 910,465 pairs fall in this category. As the genomes of more vertebrate species become sequenced, the number of these inconclusive cases should be reduced. Finally, the 168,108 pairs with score over 49 are called *confidently associated* pairs.

As shown in Figure 3–2 (B), the functional interaction score of a NCE-gene pair depends on the conservation level of its constituents. For example, large positive scores can only be reached by pairs where both the gene and the NCE are conserved back to gnathostomate ancestor. Indeed, confidently associated pairs almost exclusively involving NCEs and the genes are conserved at least as far back

bas the amniote ancestor. Pairs where either the gene or the non-coding conserved region is only conserved within eutherians generally obtain scores closer to zero.

3.4.3 Regulatory complexity

Our map of predicted gene-NCE functional interactions allows studying several aspects of gene regulation. We first classified genes based on the number of NCEs predicted to be functionally interacting with them (confidently associated pairs). We introduce here the notion of regulation complexity, a notion that is directly correlated with the number of regulatory regions regulating a gene, for example a gene regulated by many NCEs would be considered to be under complex regulation. We say that a gene has a *complex regulation* if at least 20 NCEs are predicted to interact with it, a *simple regulation* if it is linked to 1 to 5 NCEs, and a *basic regulation* if no NCE is associated to it. 619 (3.0%) genes have a basic regulation, 3921 (15.6%) genes have a simple regulation, and 2395 (11.6%) genes have a complex regulation.

Gene ontology analyses. Genes with basic and complex regulation were tested for enrichment in biological processes compared to those with simple regulation (background) using the Babelomics platform [Al-Shahrour *et al.*, 2008] (see Table 3–1). Genes with complex regulation show an enrichment for genes involved in transcription and development biological processes. Genes performing these functions (trans/dev) are known to be more evolutionary conserved than other types of biological processes. Since highly conserved genes are more likely to be associated with surrounding NCEs (see Figure 3–2, B), to control for conservation bias, we undertook the same analysis restricting the background

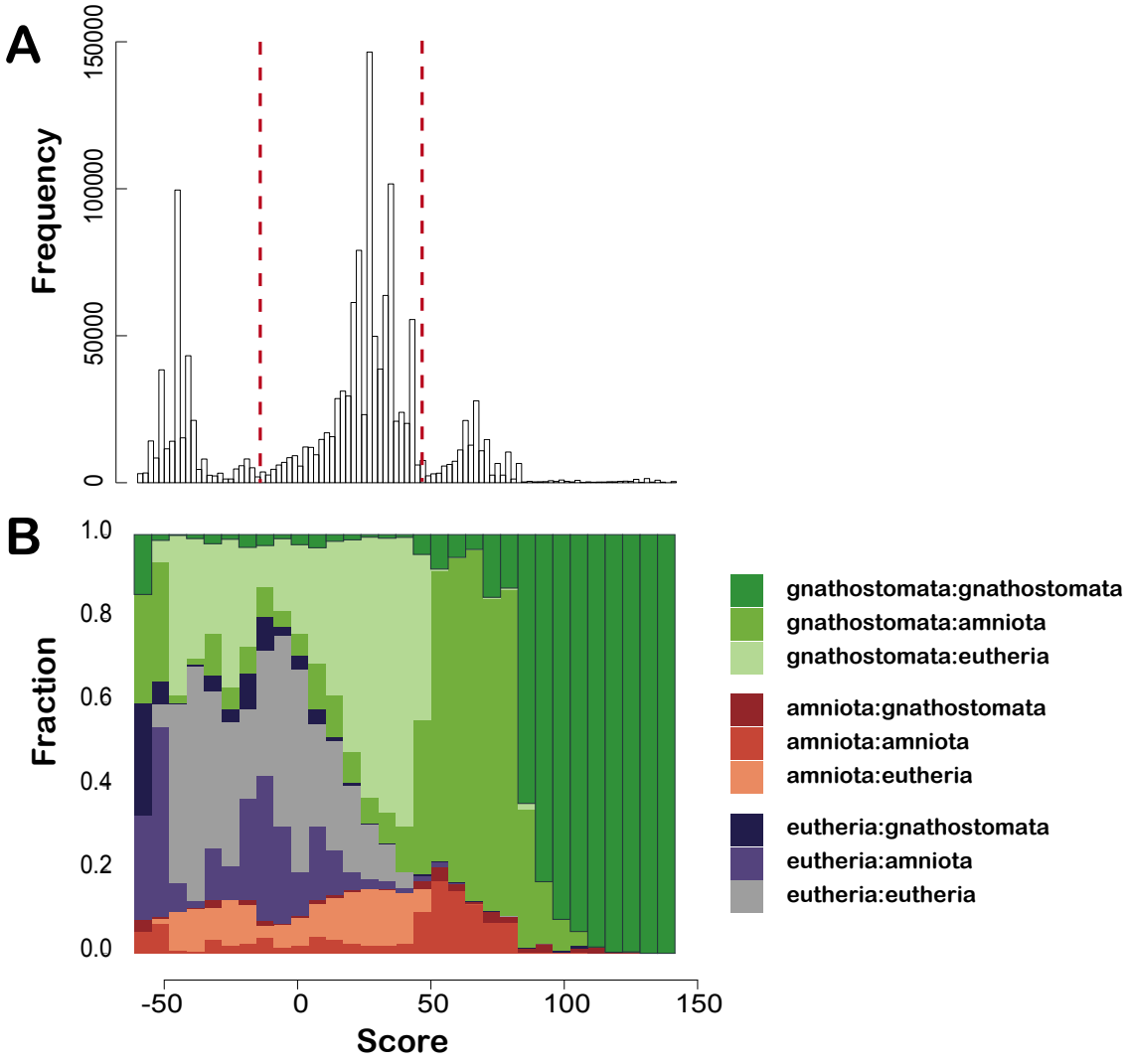

Figure 3–2: **Association score distribution.** (A) Distribution of scores for all candidate associations. (B) For each score bin, the proportion of pairs at a given conservation level is given. In the legend, the conservation is given first for the gene then for the non-coding region as follow gene:NCE.

set to only genes predating tetrapoda divergence. The results obtained show significant p-values for similar trans/dev GO categories.

Genes involved in transcription and developmental processes have complex spatio-temporal expression patterns. Such regulation maybe programmed by many *cis*-regulatory regions, which are specific to direct the expression of the gene in different tissues at different developmental times [Howard & Davidson, 2004]. Such genes are also known to be located in the vicinity of, and remain in synteny with, gene deserts [Ovcharenko *et al.*, 2005], regions known to contain a large number of *cis*-regulatory regions [Nobrega *et al.*, 2003, Ovcharenko *et al.*, 2005, de la Calle-Mustienes *et al.*, 2005].

Table 3–1: **GO analysis of highly associated genes versus background.** Level 3 GO biological processes are calculated with a double fisher-test on the Babelomics plateform. Enriched categories are sorted by decreasing enrichment.

GO category (level 3)	% in associated set	% in background	Fold increase	Corr. p-value
reproductive process (GO:0022414)	2.56	1.19	2.15	1.1e-2
multicellular organismal dev. (GO:0007275)	25.2	15.49	1.6	8.7e-13
anatomical structure dev. (GO:0048856)	21.72	15.37	1.4	4.7e-06
cellular developmental process (GO:0048869)	19.89	15.1	1.3	5.7e-4
regulation of biological process (GO:0050789)	41.06	31.55	1.3	1.2e-08
macromolecule metabolic process (GO:0043170)	54.48	47.08	1.1	5.3e-05
primary metabolic process (GO:0044238)	62.36	55.65	1.1	2.2e-04
cellular metabolic process (GO:0044237)	63.45	56.88	1.1	2.6e-4
cellular component org. and biog. (GO:0016043)	18.06	21.79	-1.2	2.3e-2
establishment of localization (GO:0051234)	17.75	23.1	-1.3	2.9e-4

When compared to the background, the genes with basic regulation show enrichment for GO biological processes involved in neurological processes and adaptive processes including "response to biotic stimulus", "defense response", "immune response" and "cell communication" (see Table 3–2). Similar enrichments have previously been observed in highly evolutionary rearranged regions [Larkin *et al.*, 2009, Mongin *et al.*, 2009]. Genes labelled to have basic regulation

are enriched for adaptive processes. These genes most probably lie within heavily rearranged regions, which explains why no associations are detected.

Table 3–2: **GO analysis of genes with basic regulation.** Level 3 GO biological processes are calculated with a double fisher test on the Babelomics plateform and corrected for multiple testing, using the genes with simple regulation as background.

GO category (level 3)	% in associated set	% in background	Fold increase	Corr. p-value
response to biotic stimulus (GO:0009607)	5.34	1.31	4.1	5.7e-05
neurological process (GO:0050877)	26.41	6.8	3.9	2.3-22
defense response (GO:0006952)	6.53	1.88	3.5	5.0e-05
immune response (GO:0006955)	8.9	2.61	3.4	2.3e-06
cell communication (GO:0007154)	38.28	27.59	1.4	4.1e-4
macromolecule metabolic process (GO:0043170)	36.8	47.08	-1.3	1.8e-3
cellular metabolic process (GO:0044237)	43.92	56.88	-1.3	5.7e-05
primary metabolic process (GO:0044238)	41.84	55.65	-1.3	2.0e-05
biosynthetic process (GO:0009058)	5.93	10.8	-1.8	1.3e-2
establishment of localization (GO:0051234)	10.98	23.1	-2.1	1.8e-06
response to stress (GO:0006950)	3.26	7.61	-2.3	7.3e-3
multicellular organismal dev. (GO:0007275)	6.53	15.49	-2.4	2.0e-05
anatomical structure dev. (GO:0048856)	5.93	15.37	-2.6	6.1e-06
cellular developmental process (GO:0048869)	5.34	15.1	-2.8	1.9e-06
cell cycle (GO:0007049)	2.67	7.92	-3.0	1.1e-3
death (GO:0016265)	1.78	5.73	-3.2	4.1e-3
regulation of biological quality (GO:0065008)	1.48	5	-3.4	7.3e-3
cell proliferation (GO:0008283)	1.48	5.42	-3.6	3.0e-3
protein localization (GO:0008104)	1.48	5.88	4.0	1.3e-3
cellular component org. and biog. (GO:0016043)	6.82	21.79	-4.1	6.0e-11

Differences in levels of regulatory complexity are also observed for different types of transcripts (protein-coding, snoRNA, miRNA) (see Table 3–5 and Figure 3–3). Small nucleolar RNAs (snoRNA) are predicted to interact with an average of only 12.4 NCEs whereas miRNAs have nearly twice as many, with 22.7 (p-value = 0.018, two-sided Wilcoxon rank sum test). Protein coding genes stand in between with a mean of 16.9. The number of other types of RNA genes was too small for this analysis.

miRNAs are short 22 nucleotides RNA molecules transcribed by RNA polymerase II [Cai *et al.*, 2004] that regulate the stability and translation of

Table 3–3: **Number of modules attached to various types of transcripts**

Transcript type	Min	Median	Mean	Max
snoRNA	1	3.5	12.4	75
coding	1	7	16.9	265
miRNA	1	8	22.7	265

mRNAs [Bushati & Cohen, 2007]. miRNAs play a key role in cellular differentiation and are tightly regulated during development (miRNA are reviewed in [Kloosterman & Plasterk, 2006, Mattick & Makunin, 2006]). Their regulation, similarly to developmental genes, is probably under the complex control of various NCEs. In contrast, snoRNAs are mainly involved in rRNA nucleotide modifications (although it seems that some show tissue specificity and developmental regulation) [Mattick & Makunin, 2006]. Therefore, snoRNA are less likely to be tightly regulated as their main role is to participate into housekeeping functions.

Genes belonging to different regulatory complexity classes have different GC content promoters. We next investigated whether the promoters of genes in each category of regulation complexity had distinguishing features. The GC content of the "promoter" (500bp region upstream of the TSS) is markedly different for each class (see Figure 3–4). Surprisingly, the average GC content of the promoters of genes with basic regulation is significantly lower than that seen for the background set (51.7% vs 58.8%; p-value=$< 2.2e^{-16}$ on a two-sided t-test). The GC content of promoters of genes with complex regulation (56.6%) is also low compared to the background set, although not as substantially (p-value = $2.3e^{-12}$).

Promoter CpG content is tightly linked to the type of genes and how they may be regulated; CpG islands are associated with housekeeping genes

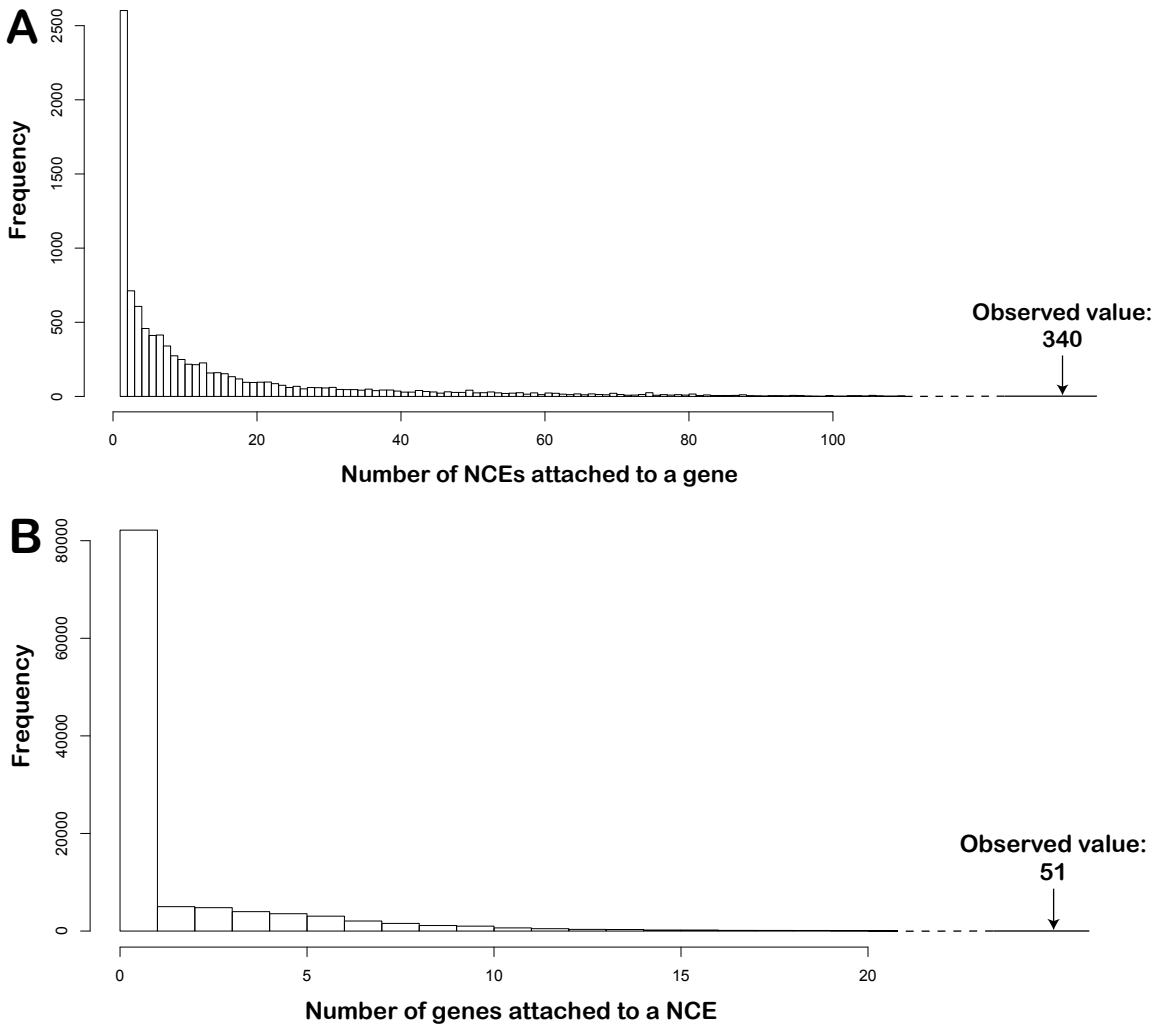

Figure 3–3: (A) Distribution of the number of NCEs predicted to interact with a gene. (B) Distribution of the number of genes predicted to interact with a NCE.

[Yamashita *et al.*, 2005, Farré *et al.*, 2007] and to genes with TATA independent transcription initiation [Carninci *et al.*, 2006]. These results may reflect that the different classes of genes we defined depending on their regulation complexity are under different regulatory processes.

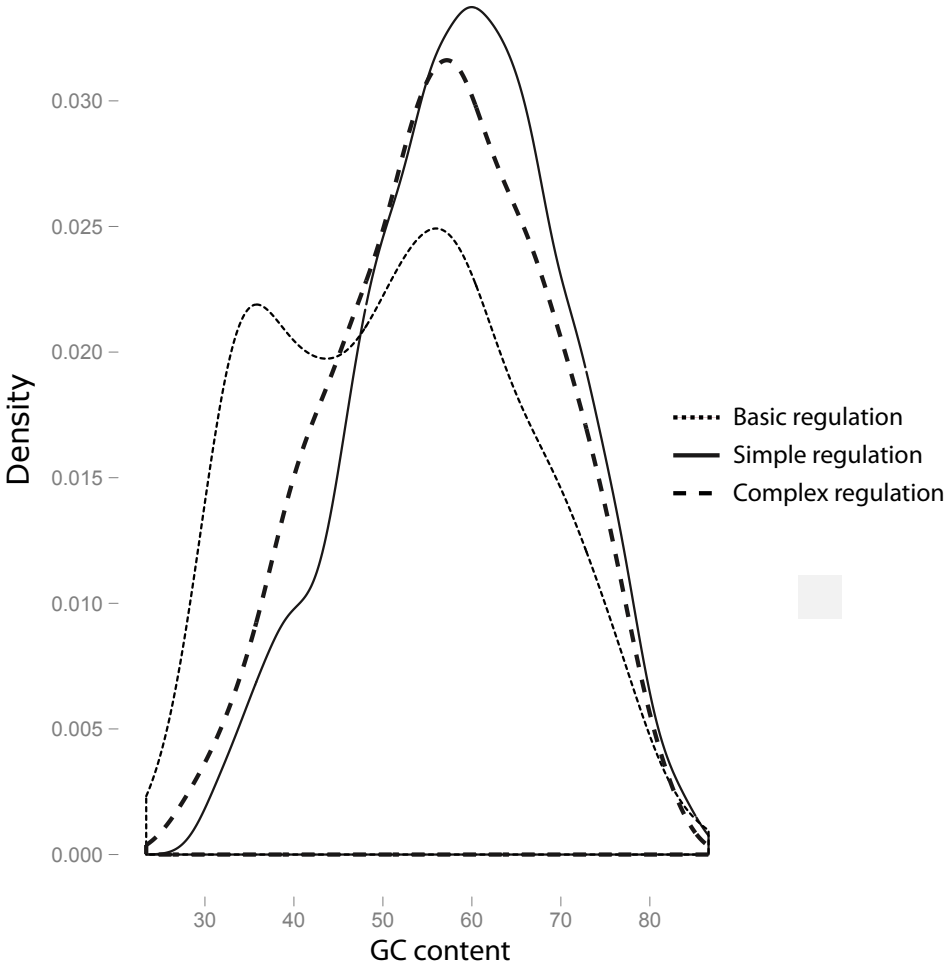

Figure 3–4: **Distribution of GC content of upstream regions for genes of each regulatory complexity class.**

Associated modules are enriched for enhancer regions. Similarly to our gene based analysis, we analyzed NCEs depending on the number of genes they were predicted to interact with. A certain fraction of NCEs have a function other that of being *cis*-regulatory regions (e.g. unannotated protein coding or RNA exons). However, one would expect that, if our interaction predictions are correct, these non-regulatory NCEs would not be predicted to interact with many other genes, except perhaps in the case of additional exons. On average, a NCE is associated to 1.5 genes, with a maximum of 51 genes (see Figure 3–3, B). We created two sets of modules depending on the number of genes with which they have been predicted to interact. The first set is composed of 4604 (3.7%) NCEs that are not predicted to interact with any genes - we call them non-interacting NCEs (niNCEs). The second set, highly interacting NCEs (hiNCEs), is composed of 3770 (3.0%) NCEs that are predicted to be functionally associated to at least 10 genes.

Different types of histone modifications have been associated to active chromatin as well as to different types of *cis*-regulatory regions [Kouzarides, 2007, Bernstein *et al.*, 2007]. We tested the overlap of both sets (hiNCE and niNCE) with genomic regions exhibiting such modifications as well as regions characterized as CTCF binding sites, as detected by Chip-Seq experiments [Mikkelsen *et al.*, 2007a].

The set of hiNCE (the most likely to be functional) has higher overlap with H3K4Me1 and H3K4Me3 docking sites (respectively corresponding to enhancer and promoter regions) than the niNCE dataset (see Table 3–4). This difference is highly significant in all cases. This result could be biased by the higher average

Table 3–4: **Overlap with histone modification data by module type.** Number of niRNA and hiRNA overlapping regions marked with different types of histone modifications. P-values are calculated with a two-sided Fisher test.

Chromatin annotation	niNCE (%)	hiNCE (%)	Fold increase	pvalue
H3K4Me1 (enhancers)	0.9	12	14	$< 2.2e^{-16}$
H3K4Me3 (promoters)	0.2	8.1	40.5	$< 2.2e^{-16}$
CTCF	0.6	2.6	4.3	$1.7e^{-13}$

level of conservation of hiNCEs compared to niNCEs. However, even when we control for this by taking into consideration only NCEs that existed predating tetrapoda divergence, the overlap with H3K4Me1 and H3K4Me3 regions remains significative with respective fold increases of 7.8 (p-value = $8.9e^{-11}$) and 30.3 (p-value = $3.4e^{-11}$). These results indicate that niNCEs and hiNCEs play different roles. Since this classification is only based on our interaction predictions, this constitutes an indirect validation of our predictions.

3.4.4 Examples

Figure 3–5 shows a clear case where *cis*-regulatory modules are linked by the predictor to genes on the right side (red links) but not on the left side where many synteny breaks in the different vertebrate genomes - which can be visualized on the net tracks - indicate that there is strong evidence that *cis*-regulatory regions and genes are not functionally linked. This non-typical example (not many regions of the genome are so unambiguous) was chosen on purpose to illustrate graphically breakage of synteny and their consequences in association predictions.

In another example, (Figure 3–6), the module is preferentially associated to a gene further apart as the closest gene is not in synteny with the NCE in most

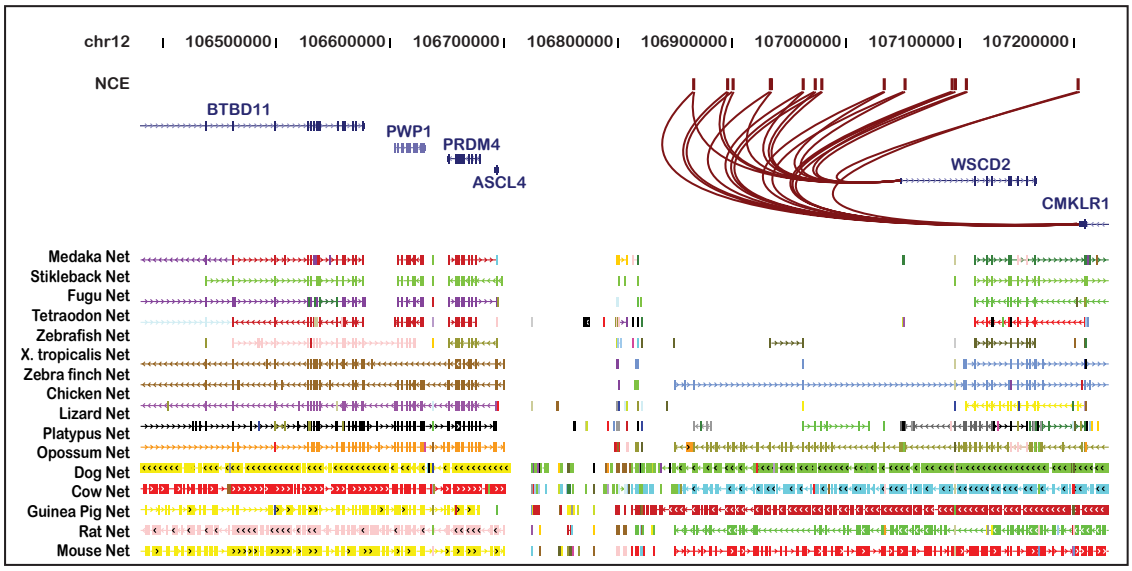

Figure 3–5: **Example of predicted interactions.** This example is chosen on purpose for its simplicity as some candidate association between the NCEs on the right side of the figures and the genes on the left side are broken in all vertebrate genomes but human. Putative association are presented in red. This figure was created and adapted from an image export from the UCSC genome browser.

genomes. This example also shows that extreme conservation is not mandatory to detect putative functional pairs.

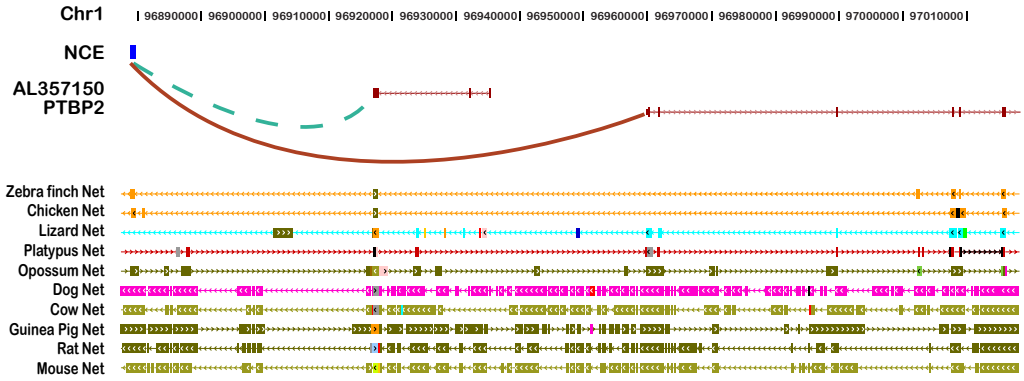

Figure 3–6: **Example where highest association score for a NCE is not the closest gene.** This example shows a case where the closest gene (AL357150) to a NCE does not get the best score. In this example, the association between the NCE and AL357150 gets a lower score (than with PTBP2) because AL357150 does not remain in synteny with the NCE. This figure was created and adapted from an image export from the UCSC genome browser.

3.5 Discussion

Linking long-range regulatory elements to the gene(s) they regulate is a challenge that remains mostly unsolved for both experimental and computational biologists. There currently exists no high-throughpout experiment that can unambiguously identify target genes for a given long-range *cis*-regulatory region.

One medium-thoughput approach is to correlate the regulatory element's activity (e.g. measured using a reporter assay) to the expression pattern of

nearby genes (e.g. measured by in-situ hybridization) to identify those whose expression domains include that of the regulatory region [Spitz *et al.*, 2003, Nobrega *et al.*, 2003, Schroeder *et al.*, 2004]. Although this approach presents some evidence of functional interaction, the results are often ambiguous. Furthermore, this approach requires data that is typically not currently available on a large scale and its predictions remain at best educated guesses.

Perhaps more promising are approaches studying high-level chromatin conformation, such as Chromatin Conformation Capture Carbon Copy (5C, [Dostie *et al.*, 2006, Fraser *et al.*, 2009]) or CHIA-PET [Fullwood *et al.*, 2009], where physical (3D) proximity between an enhancer and a promoter can be assessed in living cells. Still, these technologies remain in their infancy and the accuracy of their predictions is unknown.

Computational approaches are even more powerless to establish functional links between regulatory elements and genes, and most simply associate regulatory elements to the closest gene(s).

3.5.1 Mapping NCE/gene functional interactions is key to most gene regulation studies

In this paper, we proposed an approach based on comparative genomics to detect functional interactions between non-coding conserved elements and genes. The results of our approach is a map that associates each NCE to zero, one, or more genes, based on the conservation of their synteny in vertebrates. To our knowledge, this constitutes one of the first genome-wide maps of NCE/gene association, although efforts in this direction have already been made ([Sun *et al.*, 2008]). Despite its relatively low resolution (in most cases, the lack of divergence time

prevents us from clearly associating each NCE to a unique gene), this map will be useful to researchers in a number of areas.

We believe that our dataset is particularly useful for two types of research; laboratory research and large scale computational analysis. In the context of laboratory research, our predictions are particularly beneficial when deciphering the different regulatory elements of a gene. Putative *cis*-regulatory regions to be tested *in vivo* can be prioritized depending on their scores, high scores being tested in priority. Therefore, the researcher can save time by testing first the putative *cis*-regulatory regions that are the most likely to be be associated with the gene of interest and therefore more likely to be functional.

Large-scale *in silico* analysis of *cis*-regulatory regions often need to associate those regions to putative target genes. Most of these analysis usually associate by default to the regulatory regions closest to the genes and would greatly benefit from the association score dataset that we present here. For example, such a dataset would be useful in the context of finding target genes for *cis*-regulatory SNPs in studies such as [Ge *et al.*, 2009].

3.5.2 Predicted interactions provide insights into gene regulation

Several observations support the relative accuracy of the map produced. First, NCEs that are predicted to be interacting with several genes have the hallmark signatures of regulatory elements (histone modifications), whereas those associated to no gene do not. The latter group of NCEs likely includes non-regulatory functional regions that may be unnanotated coding or non-coding exon, matrix

attachment regions, etc. Our map is not only useful as a resource, but it also lets us learn much more gene regulation and its impact on genome evolution.

The number of NCEs associated to specific types of genes varies significantly across families. For example, miRNAs are associated to almost twice as many NCEs than are snoRNAs, and slightly more than protein-coding genes. SnoRNAs are indeed predominantly involved in housekeeping functions and are less likely to be under the influence of complex regulatory mechanisms than miRNAs. On the contrary, miRNAs are predominantly involved in developmental and cellular differentiation processes and are thus more likely to be under tight regulatory control, which would potentially require a larger number of *cis*-regulatory regions controlling the expression of the gene.

Similar observations are made across protein-coding gene families. Genes associated to a large number of NCEs are predominantly involved in transcriptional and developmental biological processes [Nelson *et al.*, 2004]. These classes of genes need to be regulated with precision at different times and in different tissues. Such specificity is achieved with an intricate set of *cis*-regulatory regions which specifically direct the expression of the gene with different spatio-temporal specificity [Howard & Davidson, 2004].

Whereas conservation of synteny is favorable for genes that are under strict regulatory constraints, there is emerging evidence that genome rearrangement hotspots [Pevzner & Tesler, 2003, Murphy *et al.*, 2005] may act as a cauldron that favor positive selection in certain regions of the genome [Larkin *et al.*, 2009]. For example, genes involved in immune response or response to stimuli - whose

adaptability is crucial - are over-represented in regions flanking evolutionary breakpoints [Larkin *et al.*, 2009, Mongin *et al.*, 2009]. These observations are consistent with our results showing an enrichment for these types of genes among genes with few predicted interactions.

Finally, our results are interesting in the context of the study of genes involved in position-effect diseases (reviewed in [Kleinjan & van Heyningen, 2005]). For example, PAX6 (aniridia [Kleinjan *et al.*, 2001]) is linked to 92 NCEs, PITX2 (Rieger syndrome [Trembath *et al.*, 2004]) to 91 NCEs, and SOX9 (Campomelic dysplasia [Bagheri-Fam *et al.*, 2001]) to 83 NCEs, whereas the average of NCE associated to a gene is 16.9. This is consistent with the dramatic phenotypic consequences observed when rearrangements occur in these regions and break the functional interaction between a gene and its regulatory regions.

3.5.3 Need for more genomes

Conservation of synteny between a NCE and gene may be due to the presence of a functional interaction between the two or to the lack of divergence time for a genome rearrangement to have separated them. At the time of the study, we were limited to the comparison of the human genome to that of 16 other vertebrates whose genomes were assembled in supercontigs of at least 10 Mb. As seen in Figure 3–2, a large number of NCE-gene pairs, especially those involving eutherian-specific NCEs, receive low-confidence scores due to the insufficient evolutionary evidence. The flip side is that old NCEs (e.g. those shared by all vertebrates) are more accurately mapped to their target gene.

The level of divergence (away from human) of the genomes considered impacts the amount of information a new sequence provides. Fish genomes have undergone a lot of rearrangements - which is good for our study -, but only 2% of human NCEs can be traced back to these species (see Table 3–5). Placental mammalian genomes share the most NCEs with human, but have typically undergone and a small number of rearrangements. Improved resolution can only be obtained by increasing the number of genomes compared, at various degrees of divergence (especially marsupials, birds, and reptiles). With whole genome sequencing becoming increasingly affordable, we expect that the accuracy of our approach will quickly increase significantly.

3.6 Methods

3.6.1 Data selection

We retrieved human non-coding conserved regions from the human 28-way [Miller *et al.*, 2007] alignment dataset available on the UCSC genome browser [Kent *et al.*, 2002], excluding any region with any overlap with EnsEMBL exons, mRNAs, or repeatMasker regions. Only regions with a score over 400 and a length over 100 are retained for further analysis. Exons were retrieved from the Ensembl (version 54) human gene prediction dataset (but excluding predictions labelled as pseudogenes)[Hubbard *et al.*, 2009].

3.6.2 Mapping

Both human non-coding conserved regions and coding regions were mapped with liftover [Kent *et al.*, 2002] (using blastz nets) to the following genomes: mouse (Build 37, july 2007) [Waterston *et al.*, 2002], rat (version 3.4, November 2004)

[Gibbs *et al.*, 2004, Havlak *et al.*, 2004], guinea pig (Broad Institute cavPor3, Feb. 2008), dog (assembly 2.0, May 2005) [Lindblad-Toh *et al.*, 2005], cow (version 4.0, October 2007)[Sequencing *et al.*, 2009], opossum (January 2006, Broad Institute), platypus (Mar. 2007, WUSTL) [Warren *et al.*, 2008], chicken (v2.1, May 2006) [Hillier *et al.*, 2004], zebra Finch (Jul. 2008, WUSTL), lizard (v1.0, Broad Institute), X. tropicalis (v4.1, DoE Joint Genome Institute), zebrafish (July 2007 , Wellcome Trust Sanger Institute), stickelback (v1.0, The Broad Institute), tetraodon (V7, Feb 2004, Genoscope), fugu (v4.0, Oct. 2004, JGI) [Aparicio *et al.*, 2002], and medaka (v1.0, Oct. 2005) [Kasahara *et al.*, 2007]. Genomes with less than 3X sequence assembly coverage were not selected since the average length of their supercontigs is usually too small (< 1 Mb) to be used in this study.

The mapping process of both coding and non-coding regions is composed of two steps. First, all human non-coding conserved regions and exons are mapped to the 16 target genomes. Second, each mapped region is mapped back to human. Only hits which map back to the same original region in human are kept (reciprocal best hits). In the case of multi-exon genes, they are considered to be mapped to a given genome if at least one of their exons is. Each human gene and non-coding conserved region is thus mapped to either zero or exactly one position in the genome of each other species. The quality of the blastz nets used for this mapping greatly limit the cases of spurious mapping.

3.6.3 Predicting functional interaction between genes and non-coding regions

Let G_S and N_S be the set of genes and non-coding conserved regions that have been mapped from human to species S. Let P_S be the set of all pairs of gene and non-coding region from species S that are located at most δ_S base pair apart (on the same chromosome) in the genome of S. Note that genes can be paired with several non-coding regions (or to none at all), and non-coding conserved regions can be paired with several genes (or to none at all). We are interested in classifying the pairs from P_{human} into functionally associated or non-functionally associated, based on the presence of the pair in the 16 other species. We set δ_{human} to 1 Mb and adjust the distance thresholds for other species in proportion to their genome size, relative to human: $\delta_s = 1.25Mb \cdot$ (GenomeSize(s)/GenomeSize(human)). Note that we use a constant of 1.25 Mb instead of 1 Mb to deal more gracefully with boundary cases where a pair may, for example, be located 0.99 Mb apart in human and 1.01 Mb apart in another specie (if both species have similar length). For any pair $(g, n) \in P_{human}$, we define the conservation status of that pair in species S as:

$$C_S(g,n) = \begin{cases} \text{conserved} & \text{if } (g,n) \in P_S \\ \text{separated} & \text{if } (n \in N_S \text{ and } g \notin G_S) \text{ or } (g,n) \notin P_S \\ \text{missing} & \text{if } n \notin N_S \end{cases}$$

3.6.4 Inference of ancestral association status

The ancestral status $C_u(g, n) \in \{\text{conserved, separated, missing}\}$ of each pair $(g, n) \in P_{human}$ is reconstructed for each ancestral node u of the phylogenetic tree,

using a variant of the Fitch algorithm [Fitch & Margoliash, 1967] for parsimonious reconstruction. When multiple ancestral statuses are equally parsimonious at the root of the tree, ties are broken in favor of separation.

3.6.5 EM algorithm

We now present two models for pairs of non-coding region and gene. The Θ^F model describes pairs that are functionally associated and for which there is selective pressure to maintain the pairing. The Θ^{NF} model describes the evolution of pairs that are not functionally associated. Each model $M \in \{F, NF\}$ is specified as follows: $\Theta^M = (P_1^M, P_2^M, ..., P_{2n-2}^M)$, where

P_u^M is the probability distribution over states node u with model M, i.e. $P_u^M(a)$ is the probability of observing state a at node u ($a \in$ {conserved, separated, missing}).

Let $A(g, n) \in$ {functional, non-functional} of each pair in $(g, n) \in P_{human}$ be the true (but unknown) functional status of pair (g, n). Because the true association status of each pair is unknown, parameters of each model are estimated in an unsupervised manner using an EM-like algorithm to find maximum likelihood estimators, based on the complete set of pairs P_{human} considered. The algorithm alternates between predicting the functional status of each pair (based on the likelihood ratio of the two models) and revising the parameters of the two models based on the predicted classification. Iterating the algorithm yields estimates for the parameters and assigns log-likelihood ratio scores to each pair in P_{human}. See below for more details.

for all $(g, n) \in P_{human}$ **do**

$A(g,n) \leftarrow F$ or NF randomly

end for

repeat

for all $M \in \{F, NF\}$ **do**

for all $a, b \in \{$conserved, separated, missing$\}$ **do**

for all $u \in V(T)$ **do**

$$P_u^M(a) = \frac{|\{(g,n):A(g,n)=M \text{ and } C_u(g,n)=a\}|+1}{|\{(g,n):A(g,n)=M\}|+3}$$

end for

end for

end for

for all $(g,n) \in P_{human}$ **do**

$$LLR(g,n) = \frac{\prod_{(u)\in V(T)} P_u^F(C_u(g,n))}{\prod_{(u)\in V(T)} P_u^{NF}(C_u(g,n))}$$

if $LLR(g,n) \geq 1$ **then** $A(G,n) \leftarrow F$

else $A(G,n) \leftarrow NF$

end for

until Convergence

3.7 Supplementary material

Table 3–5: **Genes and NCEs homology mapping.** Number of human genes and NCEs mapped to the 16 vertebrate genomes used for our analysis. Genes have been broken down depending on their gene type defined by EnsEMBL.

Species	**NCE (%)**	**Protein coding (%)**	**miRNA (%)**	**snRNA (%)**	**snoRNA (%)**	**rRNA (%)**
human	123905 (100)	21406 (100)	1664 (100)	1334 (100)	717 (100)	442 (100)
mouse	104312 (84)	17873 (83)	463 (27)	21 (1)	252 (35)	4 (0)
rat	99416 (80)	17339 (81)	402 (24)	19 (1)	241 (33)	2 (0)
guinea pig	107307 (86)	18175 (84)	547 (32)	63 (4)	277 (38)	10 (2)
dog	110094 (88)	18465 (86)	581 (34)	35 (2)	291 (40)	7 (1)
cow	107717 (86)	18098 (84)	569 (34)	75 (5)	286 (39)	7 (1)
opossum	85924 (69)	15779 (73)	297 (17)	31 (2)	174 (24)	9 (2)
platypus	67135 (54)	14692 (68)	229 (13)	20 (1)	178 (24)	4 (0)
chicken	39541 (31)	12022 (56)	136 (8)	5 (0)	93 (12)	2 (0)
zebra finch	38577 (31)	11791 (55)	163 (9)	8 (0)	91 (12)	2 (0)
lizard	31515 (25)	13254 (61)	196 (11)	46 (3)	115 (16)	53 (11)
x. tropicalis	14330 (11)	12129 (56)	137 (8)	7 (0)	73 (10)	3 (0)
zebrafish	5716 (4)	11636 (54)	140 (8)	17 (1)	59 (8)	4 (0)
fugu	3512 (2)	10173 (47)	128 (7)	25 (1)	51 (7)	5 (1)
medaka	3654 (2)	11658 (54)	125 (7)	4 (0)	45 (6)	1 (0)
stickelback	3505 (2)	10856 (50)	127 (7)	5 (0)	51 (7)	0 (0)
tetraodon	3587 (2)	11242 (52)	135 (8)	5 (0)	48 (6)	2 (0)

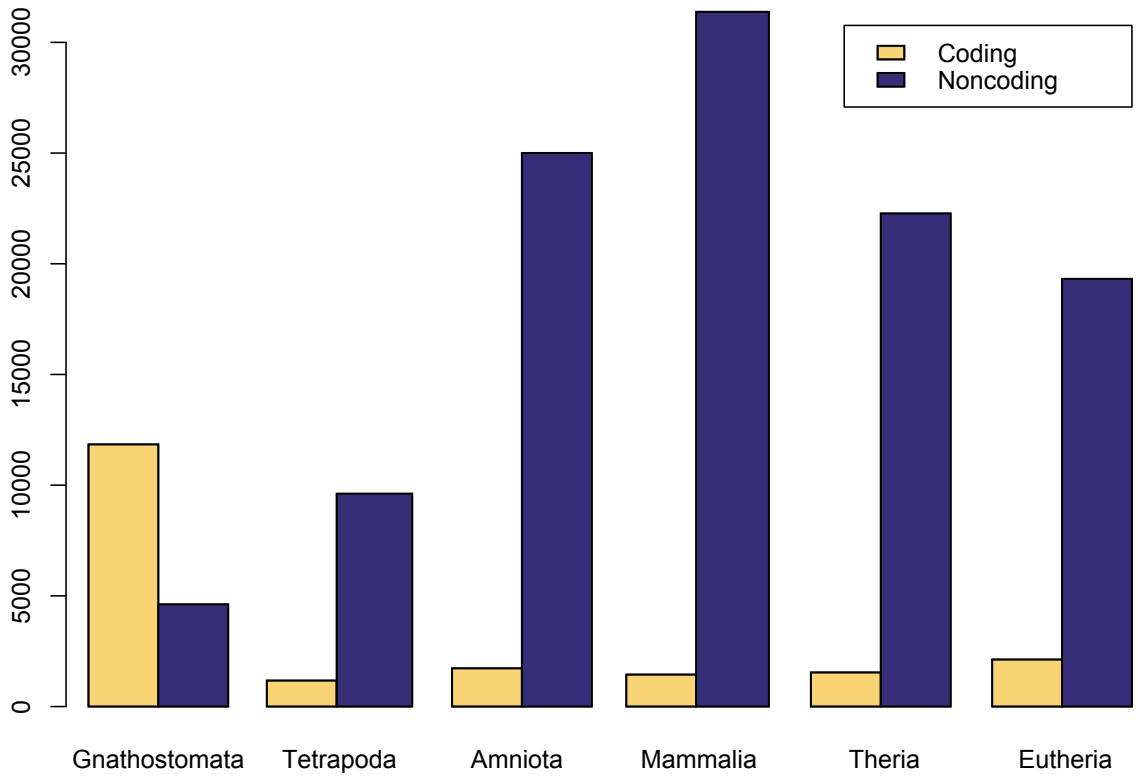

Figure 3–7: **Number of elements of each type at different conservation levels.** The conservation status of each genes and NCEs is defined as the last common ancestor of the species containing the elements.

CHAPTER 4
A new molecular tool for dissecting the developing vertebrate nervous system

4.1 Preface

In Chapter 3, we presented predictions that assess functional interaction between *cis*-regulatory regions and genes. In that context, the *in vivo* characterization of *cis*-regulatory modules is a step forward in understanding regulation processes. Therefore, we elaborated a new methodology that combines the prediction of *cis*-regulatory regions based on an adapted version of a method published by Blanchette *et al.* [Blanchette *et al.*, 2006] and a novel biomolecular method to characterize those predictions *in vivo* in medaka fish.

In this chapter, the method we use to detect *cis*-regulatory regions is more elaborate than the method we employed in the previous chapters. In Chapters 2 and 3, regulatory region predictions are based on non-coding conserved elements. In this chapter, the prediction method is based on binding site predictions, over-representation of many binding sites for a few transcription factors within a given window, and comparative genomics among teleost fish.

Medaka was chosen for three main reasons: i/ the medaka embryo is transparent so there is no need to kill it to observe the expression pattern. Therefore, the expression specificity can be monitored at different stages of development.

ii/ Medaka has a short generation time and testing regulatory regions is therefore a matter of weeks. iii/ Medaka is a particularly effective organism to create transgenic lines, feature that is important when working on expression specificity.

To develop the biomolecular method, we focussed on a subset of our predictions that show conservation with human. The reason is that putative regulatory regions conserved among vertebrates would be more likely to be functional and therefore more adapted to develop a new experimental protocol. With *in silico* analysis and observation of expression patterns from the regions tested, we realized that these datasets are enriched for regulatory regions involved in nervous tissue developmental processes. Although the original goal was to develop this combined approach to predict and characterize regulatory regions (a goal that was achieved), we decided to focus our paper on the nervous tissue specificity of our dataset. This is the work presented in this chapter and entitled "A new molecular tool for dissecting the vertebrate nervous system". This Chapter corresponds to a manuscript that has been submitted to Nature Methods as follows:

- **Emmanuel Mongin**, Thomas Auer, Frank Bourrat, Franziska Gruhl, Ken Dewar, Mathieu Blanchette, Jochen Wittbrodt, Laurence Ettwiller. A new molecular tool for dissecting the developing vertebrate nervous system. *Submitted to Nature Methods.*

4.2 Abstract

The developing vertebrate nervous system contains a remarkable array of neural cells organized into complex, evolutionarily conserved structures. The

labeling of living cells in these structures is key for the understanding of brain development and function, yet the generation of stable lines expressing reporter genes in specific spatio-temporal patterns remains a limiting step. In this study we present a fast and reliable pipeline to efficiently generate a number of stable lines expressing a reporter gene in multiple neuronal structures in the developing nervous system in fish. The pipeline combines both the accurate computational genome-wide prediction of neuronal specific *cis*-regulatory modules (CRMs) and a newly developed experimental setup to rapidly obtain transgenic lines in a cost-effective and highly reproducible manner. 95% of the CRMs tested in our experimental setup show enhancer activity in various and numerous neuronal structures belonging to all major brain subdivisions. This pipeline represents a significant step towards the dissection of embryonic neuronal development in vertebrates.

4.3 Introduction

In recent years we have witnessed a flood of new discoveries in neuroscience largely resulting from the ability to monitor and specifically manipulate living cells in the context of the developing nervous system using reporter gene expression [Tsien, 1998]. The resulting transgenic lines expressing reporter genes in a time and space restricted manner have been a breakthrough in modern biology. Recently, exciting developments in engineering new proteins has extended current barriers to allow monitoring and manipulating the activity of specific pathways within living cells [ichi Higashijima *et al.*, 2003, Nagai *et al.*, 2001,

Pertz *et al.*, 2006, Srivastava *et al.*, 2007]. Nonetheless, these techniques rely heavily on the ability to stably drive gene expression to specific developmental stages, brain structures and cell types. While great efforts have been made to facilitate the obtention of such stable lines, this step remains a serious bottleneck.

The most widely used strategy to express reporter genes in anatomical structures relies on the use of regulatory elements, often promoters of genes known to be expressed in the desired structures. This trial and error process is slow and tedious. Thus to maximize the chances of getting the right regulatory sequences, other approaches have used entire loci around selected genes employing BAC technology [Heintz, 2001]. However, this methodology is time-consuming and the level of reporter expression may not be high enough for proper monitoring. Other attempt to generate reporter gene expression in various structures are based on the random insertion of a reporter cassette in fish genome. In some cases the construct is activated by nearby regulatory element(s) resulting in the selective activation of the reporter gene [Parinov *et al.*, 2004, Ellingsen *et al.*, 2005, Korzh, 2007].

Despite advantages of one approach over another, all these methodologies aimed at generating stable transgenic animals have the significant drawback to lack specificity. Testing semi-random elements either by promoter bashing or enhancer traps results in a very low success rate, while BAC technology, which addresses the specificity issue by using the entire locus instead, is experimentally costly and cannot be easily scaled up.

In parallel, progress has been made towards the computational identification of regulatory regions in sequenced genomes. Previous work has

shown that, without experimental priors, functional constraints acting on non-coding sequences are one of the most predictive information to locate regulatory elements [Dermitzakis *et al.*, 2002, Bejerano *et al.*, 2004b]. Thus cross-species comparison has been extensively used to improve the detection of functional non-coding DNA regions from neutrally evolving DNA [Loots *et al.*, 2000]. The discovery of new regulatory regions using inter-species conservation was greatly stimulated by the recent availability of various vertebrate genomes, from mammals to fish [Lander *et al.*, 2001, Waterston *et al.*, 2002, Hillier *et al.*, 2004, Aparicio *et al.*, 2002] as well as the development of more specific and sensitive alignment programs [Brudno *et al.*, 2003a, Schwartz *et al.*, 2003, Blanchette *et al.*, 2004, Paten *et al.*, 2009]. Furthermore, it has been shown that the tendency of transcription factor binding sites (TFBS) to cluster together can be used to predict putative *cis*-regulatory modules (CRMs) [Howard & Davidson, 2004]. This led to the development of new methods to locate clusters of binding sites in conserved regions [Philippakis *et al.*, 2005]. An algorithm that combines both, inter-species binding site conservation and clustering has recently been applied to the human genome [Blanchette *et al.*, 2006] resulting in the identification of 118,000 predicted human regulatory elements [Ferretti *et al.*, 2007].

Here, we report the development of a new pipeline aimed at specifically labeling, in a stable manner, various neuronal structures in developing *Oryzias latipes* (medaka) embryos. This pipeline represents two major breakthroughs compared to previous methodologies: a selective step to predict neuronal specific

regulatory regions, combined with a new reliable enhancer assay to efficiently obtain stable lines expressing the reporter gene in neuronal structures.

The selective step applies a modified version of the computational pipeline previously described [Blanchette *et al.*, 2006] to select a large number of short (100-1000bp) regions predicted to be regulatory in fish. In order to increase the likelihood of identifying CRMs, we further filtered this dataset keeping only 491 elements that exhibit detectable sequence conservation across vertebrates. We tested several of these regions in our new enhancer assay in the medaka fish. In this analysis, a vast majority of the regions tested resulted in a strong, reproducible expression of the reporter gene in various neuronal structures. All the major subdivisions of the medaka CNS are covered by at least one expression pattern. In most of the cases, the reporter gene expression persists beyond hatching and in all cases analyzed, at least two independent stable lines were obtained. We also showed that the enhancer activity is reminiscent of the endogenous target gene expression, which facilitates the additional selection of regions to target specific anatomical areas. Both the computational prediction of CRMs and the experimental results have been integrated into databases for easy access and queries.

The success rate in terms of stable transgenic lines expressing reporter gene in neuronal tissues is higher than previously achieved using other approaches. Thus our pipeline is an important tool for labeling neuronal structures and deciphering the regulatory grammar controlling the development of the neuronal system in vertebrates. Furthermore our results have demonstrated that pan-vertebrate

conserved non-coding elements show preferred activity in neuronal structures compared to less deeply conserved elements.

4.4 Results

4.4.1 Identification of a set of neuronal regulatory elements

One of the key steps for the establishment of a robust pipeline for the labeling of developmental structures is the accurate prediction of autonomous regulatory elements in the genome. Thus to define genomic regions most likely involved in gene regulation, we use a variant of the PreMod algorithm [Blanchette *et al.*, 2006] applied for the medaka genome (see Methods). The algorithm first identifies individual transcription factor binding sites based on a set of 402 high quality position-weight matrices (PWM), including manually curated databases of known TFBS (Transfac [Matys *et al.*, 2006], Jaspar [Vlieghe *et al.*, 2005]) and results from ChIP data [Ettwiller *et al.*, 2007]. It then assesses conservation of the predicted TFBS by comparing medaka sequence to the ortholog sequences in *Tetraodon nigroviridis* (tetraodon), *Takifugu rubripes* (puffersh), and *Gasterosteus aculeatus* (stickleback). Finally, clusters of conserved homotypic binding sites (or oligotypic composed of binding sites for up to 5 different factors) were identified and reported as candidate teleost CRMs.

The CRM prediction algorithm resulted in the identification of 23 011 short elements (average length 244 bp; median length 136 bp) which are composed in average of 62 putative transcription factor binding sites. These regions are broadly distributed across the genome, with 65% being more than 10 kb away from any annotated transcription start site (TSS) (average distance to closest

TSS: 32 kb) (Supplementary Figure 4–5). It has previously been shown that many ultra-conserved non-coding elements conserved across vertebrates are functional enhancers [Pennacchio *et al.*, 2006]. These elements are also known to be preferentially located around developmental genes and are consequently hypothesized to be active during development [Bejerano *et al.*, 2004b]. We thus selected predicted teleost CRMs for which a statistically significant alignments in a conserved syntenic block with human can be found (see Methods for details). The resulting 491 highly conserved vertebrate CRMs, while broadly distributed across the genome, are significantly found more often within 10 kb of the nearest TSS.

Both sets of predicted CRMs (teleost and vertebrate conserved CRMs) are stored in the PreMod database [Ferretti *et al.*, 2007] (http://premod.mcb.mcgill.ca). PreMod provides the location, score, and binding site content for each predicted CRM. It also reports which transcription factor matrices were used to build the CRM (tag matrices). Predicted CRMs are displayed in their genomic context and surrounding genes are identified. Where in-situ expression of medaka genes or CRM activity information is available, PreMod links to the corresponding experimental data stored in the 4DXpress database [Haudry *et al.*, 2008] (http://4dx.embl.de/4DXembl). Figure 4–1 summarizes the *in silico* and experimental procedure.

Next, we took advantage of the large compendium of *Danio rerio* (zebrafish) in-situ annotations from ZFIN [Sprague *et al.*, 2006] to shed light on the putative function of the predicted CRMs. We first mapped the in-situ annotation of the zebrafish genes onto their orthologs in medaka (See Methods). For each of those

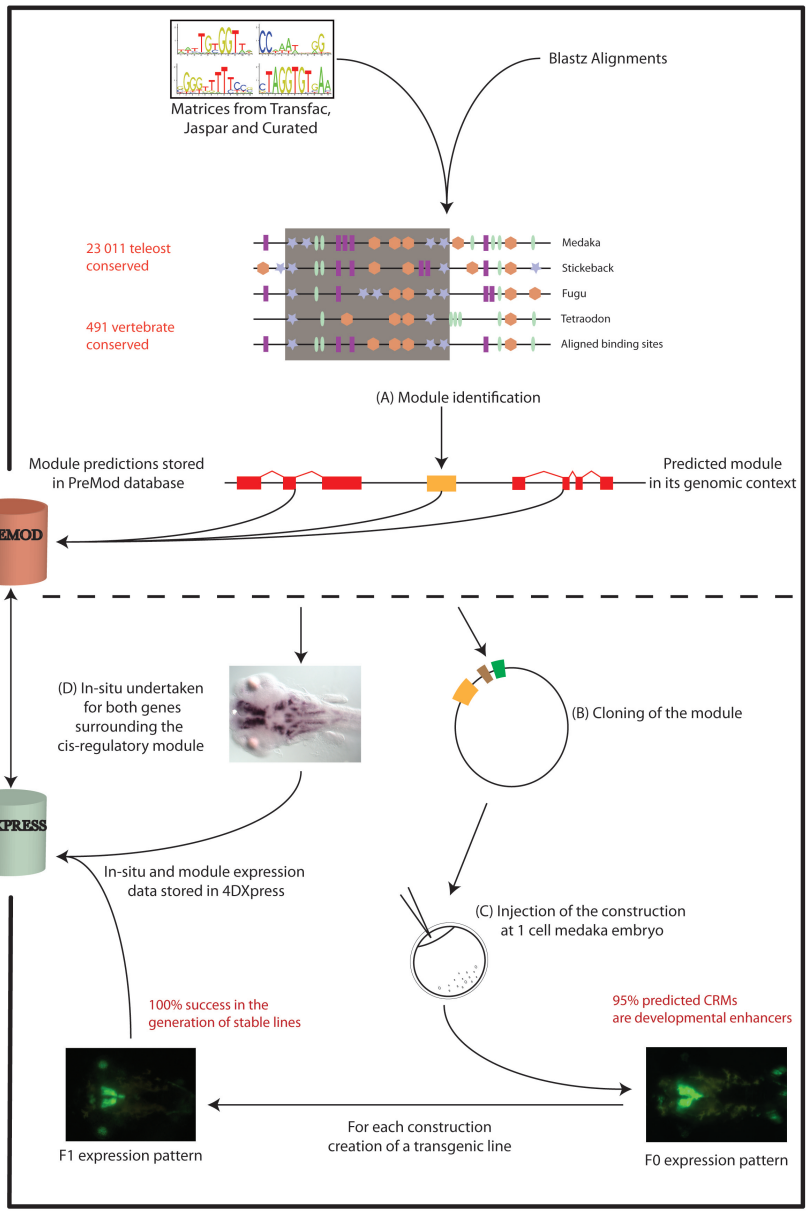

Figure 4–1: **Schema of the pipeline.** Step A is the computational predictionsof CRMs. A subset of these CRMs are then experimentally tested *in vivo* (step B, C). The expression of the flanking genes are also analysed by whole mount in-situ hybridisation (step D) and the results of the pipeline are stored in two databases (PreMod and 4DXPress).

predicted CRMs in the medaka genome, we located the closest of the two flanking genes and assigned its annotation to the CRM. We then tested if vertebrate conserved CRMs show a statistically significant increase in annotation for certain developmental tissues compared to the rest of predicted CRMs. Interestingly, we find that vertebrate CRMs are associated with an elevated ratio of genes expressed in various brain regions compared to the rest of the dataset (Figure 4–2; Supplementary Table 4–1). More specifically, 74% of vertebrate conserved CRMs are associated with genes annotated as expressed in central nervous system such as the brain, p-value = $5e^{-4}$ and spinal cord, p-value = $2e^{-3}$. On the other hand, enrichment is not observed in non-neuronal tissues (pronephros: p-value = 0.22, somite: p-value = 0.45, cardiovascular system: p-value = 0.67).

This finding, empirically observed in mouse enhancer analysis [Pennacchio *et al.*, 2006] and consistent with this study, has important implications in terms of evolution of the neuronal system in vertebrates and can be used as criteria for prioritization of regulatory elements to test when labeling of neuronal structures is pursued.

4.4.2 Development of a new enhancer assay in medaka

We developed a new enhancer assay to rapidly test genomic regions for enhancer activity and derive stable transgenic lines. In view of setting up a pipeline for large-scale analysis, we particularly focused on generating a quick and reliable readouts, which required live monitoring of the expression pattern directly in injected embryos. The ability to record GFP expression in a live embryo throughout its development is a clear advantage of the fish system compared to mouse embryo. Thus we expect an increased sensitivity in the detection of

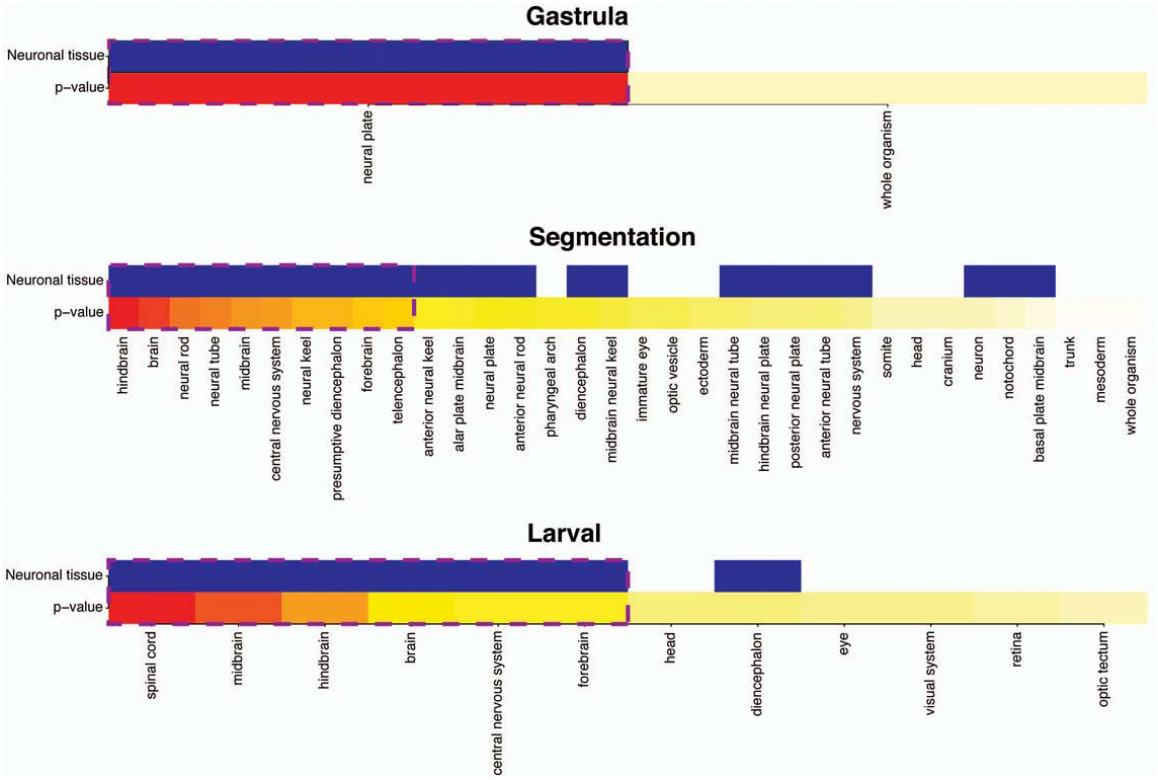

Figure 4–2: **Enrichment of vertebrate conserved CRMs around genes express in neuronal tissues.** Blue squares correspond to neuronal structures. P-values are shown with a color code, the most significant enrichments correspond to the p-values in red, the least significant to p-values in white. Significant p-value cutoff has been determined for a 5% false discovery rate (Benjamini, Hochberg method) and identified with a purple dashed rectangle. See Supplementary Table 1 for numerical values.

expression patterns and better characterization of this expression pattern through time.

We use the meganuclease mediated transgenesis [Grabher & Wittbrodt, 2007] as a method of choice to obtain a highly efficient integration of the transgene into the genome and to consequently achieve high rates of germline transmission. Predicted CRMs were first cloned into a pBlueScript-based transgenesis vector containing two recognition sites for the meganuclease ISce-I [Monteilhet *et al.*, 1990] flanking the core promoter Hsp70::gfp and a SV40-polyadenylation signal. Injected embryos were visually monitored every day for seven days to follow the spatio-temporal pattern of GFP expression during embryonic developmental stages.

The development of a robust and efficient experimental pipeline requires the ability to distinguish between the absence of enhancer activity and the failure of the injection experiment. For this, we use the convenient characteristic of the hsp70 core promoter to trigger a strong and specific lens expression beginning of stage 28 on [Blechinger *et al.*, 2002]. The heat-inducible zebrafish hsp70 gene is expressed during normal lens development under non-stress conditions. This feature remains when CRMs are cloned upstream of the core promoter, resulting in embryos with composite expression in the lens and other domain(s) (if any) specific to the CRM. As the correlation between lens expression and expression in other domains is very high when testing positive CRMs, the monitoring of lens expression itself is a very good indicator of the injection success rate.

We therefore monitored the number of lens-positive embryos (injection success rate) and the number of embryos showing reproducible GFP expression in other

domains (Supplementary Table 4–2). The percentage of successfully injected embryos showing reproducible expression outside the lens is then calculated and when above 50% (out of at least 20 successfully embryos) was used to call a genomic region positive for enhancer activity. To be considered, a consistent pattern was observable in at least 10 individual fish. This typically requires injecting less than a hundred embryos, which is easily achievable in one injection experiment. About 1 in every 50 successfully injected embryos shows non-consistent expression most likely resulting from the activity of a local enhancer (enhancer trap). In our paradigm, the enhancer trap expression pattern does not pass the quality control and is therefore discarded. This quality control measurement is a significant improvement over previously described CRM assays from which the distinction between injection failure and lack of enhancer activity cannot be made.

In typical experiments we obtained an injection success rate of 40%, and, in the case of functional CRMs, an average of 65% showed consistent expression patterns (Supplementary Table 4–2). These highly reproducible patterns are a good indication that the expression patterns we observe are solely the result of the tested CRM activity.

The efficiency of integration of the construct into the genome at early stages of development dictates the degree of mosaicism in transient lines. This aspect is particularly important in the detection of those CRMs triggering expression in only a limited number of cells. In this case, it is key to limit the degree of mosaicism to a minimum. We thus assessed the ability of our assay to reliably identify GFP expression restricted to a handful of cells only. In the case of a high

degree of mosaicism, the detection of embryos showing specific expression in a restricted amount of cell is more difficult.

One of our tested constructs (MEDMOD062451, stage 26) shows expression in a limited number of cells located bilaterally in the diencephalon (Supplementary Figure 4–6). In the resulting stable lines (stage 26) we reproducibly estimated a total of around 15 cells labeled on each sides. In transient experiments, despite the limited amount of cell labeled, we found that the majority of the injected fish have both cell populations labeled (Supplementary Figure 4–6, A-H). We also found no visible change in terms of percentage of embryos showing specific expression compared to CRMs resulting in a broader GFP expression. Additionally, all positive experiments resulted in stable lines derived from at least two independent fish. Taken together, these results strongly indicate that the construct is efficiently and early on integrated into the genome limiting de facto the mosaicism of the reporter gene expression.

Furthermore the same spatio-temporal structures are labeled in transient injected fish compare to stable lines (Supplementary Figure 4–7) showing that the accurate description of enhancer activity can be done directly on the injected fish Thus, the required experimental time can be cut back from eight weeks (generation time of medaka) to less than a week (time for embryogenesis in medaka).

In conclusion, our assay further simplifies and shortens the measurement of enhancer activity and allows the detection of numerous regulatory regions that otherwise would be called negative.

4.4.3 A vast majority of the computationally predicted regions have enhancer activity

The top ten computationally predicted vertebrate CRMs located in eight genomic loci were experimentally tested for enhancer activity and the injected fish were kept as stable lines (Supplementary Figure 4–7). To evaluate the global success rate of the pipeline, an additional ten predicted CRMs evenly distributed among the 200 top scoring candidates were tested for enhancer activity.

To ensure the inclusion of all the necessary regulatory features we fused close-by predicted CRMs (see Methods) and extended the prediction to include 200 bp flanking sequences on each side. The resulting regions ranged from about 500 bp to 2 kb and their location varies from 2095 bp to 63 755 bp away from the TSS of the nearest gene (20 kb on average)

Out of the 20 tested regions, 19 triggered a reproducible expression pattern in transient transgenic fish (Figure 4–3, Figure 4–4, and Supplementary Figure 4–8). Extrapolated to the full dataset of the 200 top scoring regions, we estimate that 95% of the computationally predicted CRMs have enhancer activity during embryonic development. This success rate is significantly higher than for another large-scale study done in mouse, which revealed that 40% of ultra-conserved elements show enhancer activity [Pennacchio *et al.*, 2006]. This higher success rate is further discussed but likely due to both the prediction method involving highly conserved regions and the monitoring of reporter gene expression throughout the whole embryonic development.

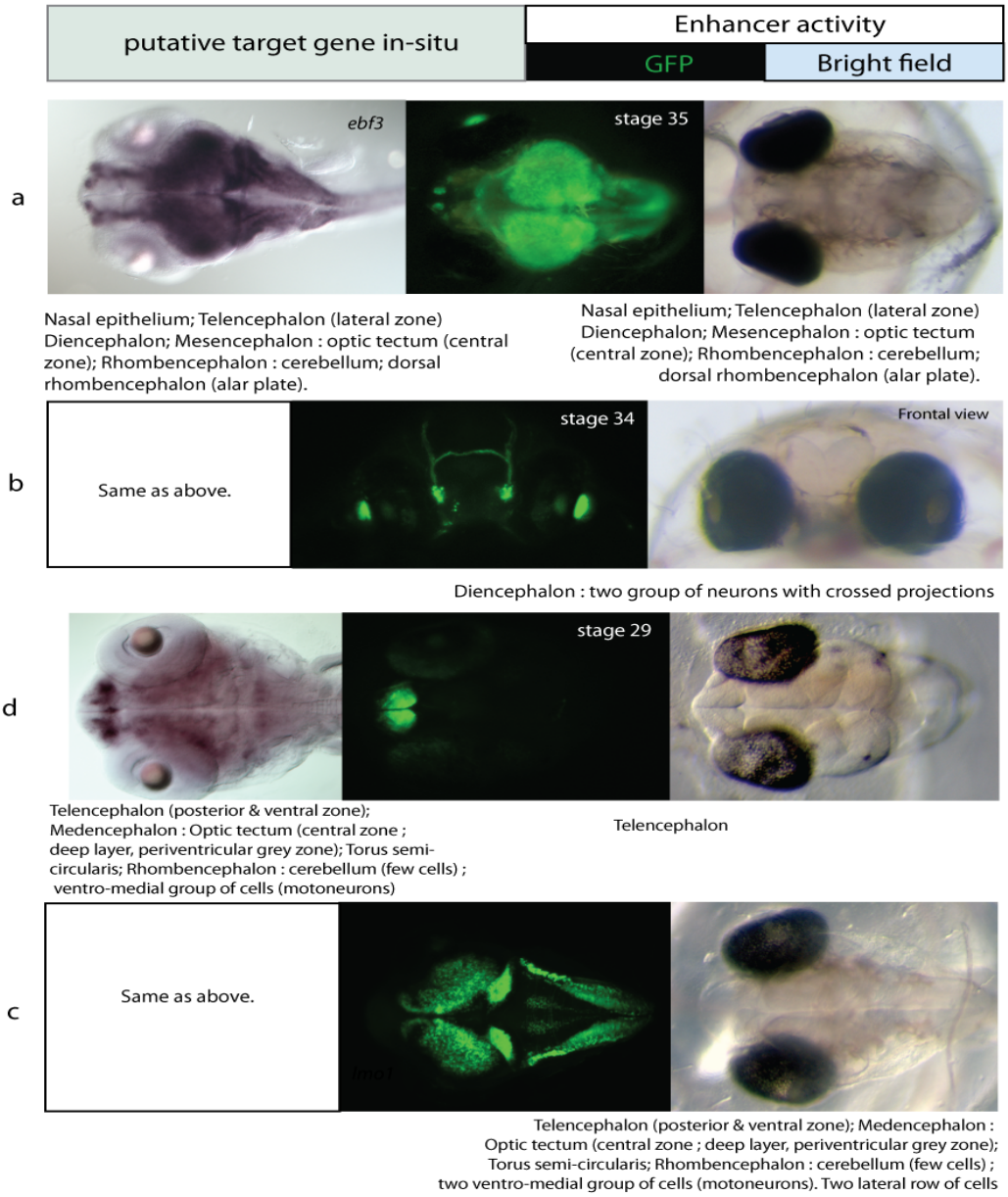

Figure 4–3: **Summary of the experimental analysis (a-d).** In-situ of the flanking gene and stable lines expressing GFP under the control of the corresponding module. MEDMOD046561 did not show detectable enhancer activity.

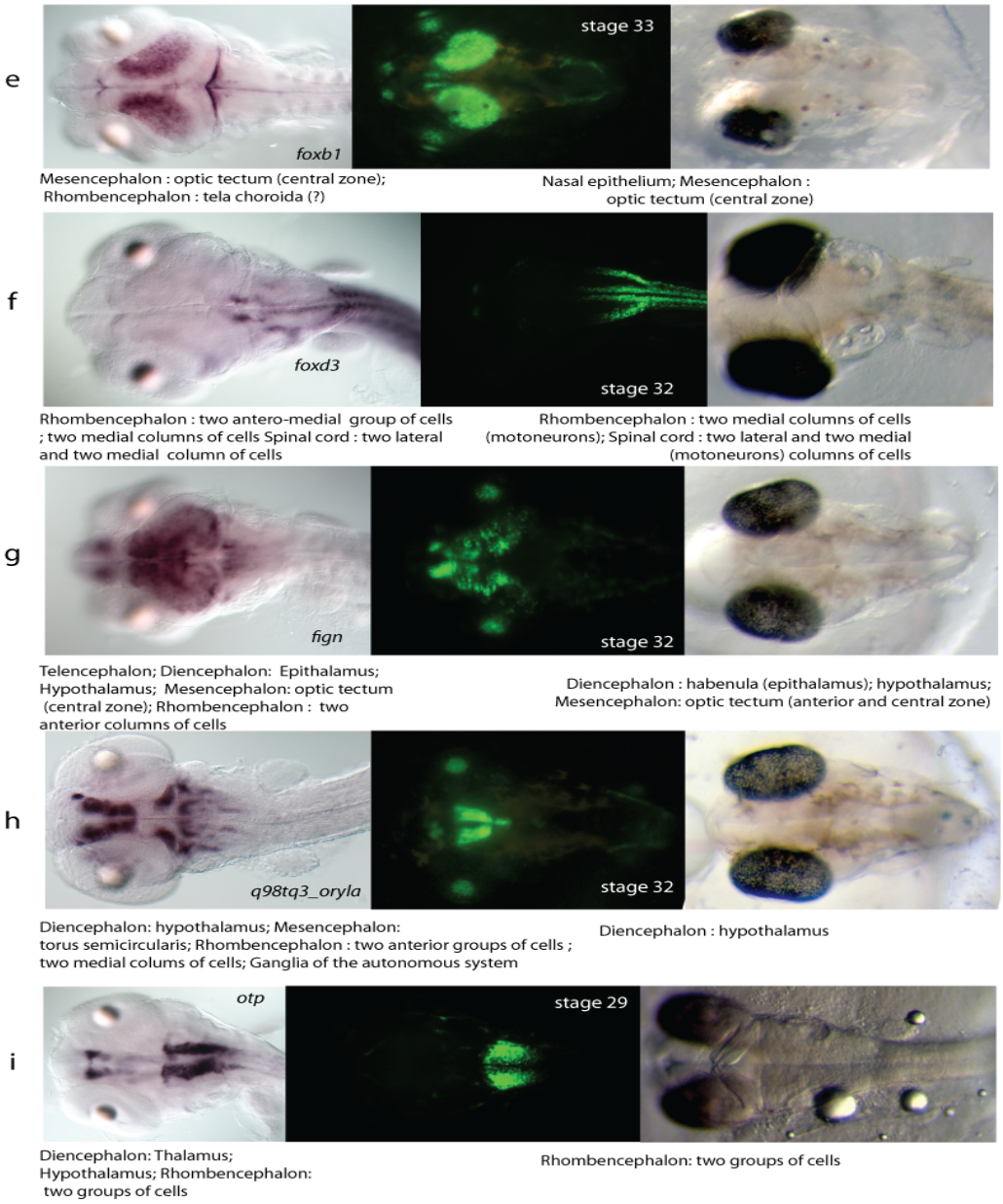

Figure 4–4: **Summary of the experimental analysis (e-i).**

4.4.4 Stable expression of the reporter genes in neuronal structures

Further confirming the computational predictions, all the positive elements drive reporter gene expression in various neuronal structures, with some limited to very specific areas of the brain or peripheral nervous system, sometimes with just a few cells being labeled. For example, MEDMOD021885 highlights a cluster of a few dozen neurons located bilaterally in the diencephalon (Figure 4–3, D). Other sequences gave broader expression patterns covering entire brain structure(s). Expression can appear spotted (even in stable lines), suggesting that only one or a few cell types are labeled (for example line MEDMOD062451, Figure 4–3, B).

For a general analysis of the neuronal system, a complete coverage in terms of labeled structures would be desirable. All major subdivisions of the vertebrate CNS have been found to include labeled cells in our assays. Thus, labeling is found in telencephalic domains (for example, line MEDMOD021953), diencephalon (lines MEDMOD021953, MEDMOD021885, MEDMOD046007), mesencephalon (lines MEDMOD074008, MEDMOD021953 for instance), rhombencephalon (lines MEDMOD021953 and MEDMOD070042, among others), and spinal cord (line MEDMOD070042) as well as in other neuron-containing structures, such as the nasal epithelium (line MEDMOD21953 and MEDMOD074008) (Figure 4–3, Figure 4–4 and Supplementary Figure 4–8).

We further proceed to a more detailed analysis of the top 10 candidate dataset composed of 9 regions with validated enhancer activity. Stable transgenic lines were generated in all experiments and these lines have been annotated using a controlled vocabulary from the medaka anatomical ontology [Smith *et al.*, 2007]

and incorporated in 4DXpress. From the 32 defined neuronal structures in the ontology, 20 (62%) were labeled in at least one of the stable lines obtained. Some structures are labeled in more than one line : for example the mesencephalic optic tectum is labeled in the MEDMOD021953, MEDMOD062451, MEDMOD074008, MEDMOD046007 lines (Figure 4–3 A,B,C and Figure 4–4 F) but overall, the spectrum of neuronal structures labeled is broad.

These stable lines expressing reporter gene in such restricted number of cell types are important starting point for further functional analysis of defined brain structures. In the long term, they offer a valuable resource for the accurate characterization of neuronal cell types and the anatomical description of the embryonic neural structure in vertebrates.

We next investigated whether the enhancer activity of the cloned fragment tested in our assay represents an accurate description of the regulatory activity of the sequence in its native endogenous locus.

First we examined the effect of the interaction between the core promoter and the CRM on the reporter gene expression. For two CRMs, we compared GFP expression pattern when using either the endogenous promoter of the corresponding putative target gene (identified based on in-situ expression, see below) or the hsp70 promoter. For one CRM (module MEDMOD086628), we observed no differences in the spatio-temporal expression of the reporter gene (data not shown). For the other CRM (module MEDMOD062451), a slight difference in GFP expression could be observed : The hsp70 promoter induces an

extension of the rhombic lips expression at stage 33 (Supplementary Figure 4–9). With both endogenous promoters the overall intensity of the pattern seems slightly reduced compared to hsp70. Nevertheless these results are encouraging and providing evidence that the hsp70 core promoter itself does not significantly alter the expression specificity of the CRM.

Next, we investigated whether the reporter gene expression monitored in our stable lines reflects the expression pattern of the genes surrounding the CRMs in their native genomic location. For this we performed whole-mount in-situ hybridization on the genes flanking the CRM regions and compared the resulting expression patterns with the activity of the enhancers (Figures 4–3 and 4–4). For each of the 9 predicted CRMs showing enhancer activity, we found that at least one of the flanking genes is expressed during development. Furthermore, at least one spatio-temporal domain of expression is common to the reporter gene expression under the control of the corresponding enhancer.

Taken together, these results strongly suggest that our enhancer assay outputs represent an accurate description of the activity of the enhancers in their native endogenous state and are not the result of an artifact of the enhancer assay. This point is secondary in view of systematically labeling neuronal structures but important to consider for the investigation of the mechanisms of gene regulation. Consequently, using gene expression information, the regions to test can be further narrowed down to a subset that is likely to result in the labeling of desired structure.

4.5 Discussion

We describe a new hybrid methodology aimed at identifying neuronal regulatory elements in fish. With 95% success rate after experimental validation and a 100% success in transgenesis, this pipeline is, to date, the most efficient procedure to obtain stable transgenic lines expressing reporter genes in various neuronal structures. Furthermore, the orthologs of three of the twenty CRMs tested in this study had previously been tested in mouse [Pennacchio *et al.*, 2006]. For one of the sequences assayed (homologous to MEDMOD021953), expression of the reporter gene localized in the hindbrain of mouse at stage E11.5. In comparison, module MEDMOD021953 also shows expression in the medaka hindbrain but is not restricted to this structure. No expression was obtained for the other mouse sequence assayed (homologous to MEDMOD086628) while it drives reporter gene expression in rhombomers in our study. These results indicate the high sensitivity of the enhancer assay in medaka.

One striking result of this pipeline is the remarkable specificity observed for some expression patterns with, in some cases, only few neuronal cells labeled. Such specific expression patterns have not been described to date when testing regulatory regions in a systematic analysis. We hypothesize that such precise expression can be explained by (i) The CRM prediction method: While most of the previous methods to identify regulatory regions are solely based on inter-species conservation, the method we propose here also takes in account the over-representation of binding sites for a number of transcription factors within a certain window. We therefore believe that this approach allows for better functional specificity of the

predicted regulatory regions. (ii) The limited mosaicism allows the identification of expression patterns localized in very restricted structures. This is crucial to confidently evaluate expression pattern where the tested regulatory module is so specific that only a few cells are labeled. While striking specificity can be seen for certain lines, others show broader expression resulting in the labeling of various and numerous neuronal structures belonging to all major brain subdivisions.

We have also shown that the patterns of reporter gene expression in our lines are overlap of the expression of genes originally located in the vicinity of the tested regulatory elements. Using gene expression information such as in-situ data, it will therefore be possible to further target the pipeline to select regions most likely active in specific neuronal structures. This task is facilitated by the fact that the computational predictions stored in PreMod are linked to the in-situ stored in 4DXpress. Furthermore PreMod provides CRMs in their genomic context as well as a score for each predicted regulatory region. As a result, prior to *in vivo* testing, CRMs can be targeted in a pertinent order based on their genomic context and score.

Our pipeline, designed to create neuronal tissue specific markers, is also of great interest for researchers focusing on deciphering the different regulatory elements of a gene, to identify markers for specific tissues in development and finally as a cost effective screening tool prior to more time consuming and expensive experiments.

Finally we have shown, both experimentally and by computational prediction that pan-vertebrate conserved CRMs have preferred activity in neuronal structures

compared to less deeply conserved CRMs. This conservation may reflect the ancestrality of the structures and the functions of neuronal tissues. More in-depth analysis on such conservation can shed light on evolutionary events that leads to morphological innovation via the emergence of new regulatory interactions.

4.6 Methods

4.6.1 CRM prediction

We collected a comprehensive set of 402 non-redundant position weight matrices (PWM) based on Transfac (version 9.2) [Matys *et al.*, 2006], Jaspar core vertebrate matrices [Vlieghe *et al.*, 2005] and a curated set of matrices built from Chip data with Trawler [Ettwiller *et al.*, 2007]. Transfac matrices were ltered out based on the following rules: (i) All non-vertebrate transfac matrices were removed, except for 8 hand-picked Drosophila matrices for factors known to be involved in vertebrate development; (ii) Matrices linked to more than two different TFs (from the same species) were discarded; (iii) Among different matrices for the same TF, only that with the highest quality value was kept or, if not available, that whose predicted sites are the most conserved through vertebrate evolution (M. Blanchette, unpublished).

For each TF, binding sites were predicted in the complete non-coding, non-repetitive regions (based on Ensembl database version 41 [Flicek *et al.*, 2008] of the genomes of medaka (Oryzias latipes, assembly HdrR, Oct 2005 [Kasahara *et al.*, 2007], tetraodon (*Tetraodon nigroviridis*, assembly, Tetraodon 7, Apr 2003 [Jaillon *et al.*, 2004]), stickelback (Gasterosteus aculeatus, assembly Broad S1, Feb 2006, Broad Institute), and Fugu (*Takifugu rubripes*, assembly 1.0, Aug 2002 [Aparicio *et al.*, 2002]).

We followed the procedure described in [Blanchette *et al.*, 2006], with the following slight modications: (i) The local GC-content background model used in [Blanchette *et al.*, 2006] was replaced by a uniform background model; (ii) Interspecies binding site conservation was measured using a more exible approach that allows for (but penalizes) sites that are slightly misaligned, up to 20 bp. In addition, conservation was weighted as follows: hitScorealn(m, p)= hitScoreMedaka+max(hitScoreTetraodon, hitScoreStickleback, hitScoreFugu). hitScore will then depend on both the score of the binding site in medaka and its conservation in at least one other teleost. Note that a binding site can have a high score without being conserved if the medaka scoring hit is strong enough. CRMs are predicted genome-wide and are not targeted to specific regions (regions with known developmental genes for example).

A subset of 20 teleost CRM predictions was selected for our *in vivo* characterization, using a criterion combining high module score and conservation with human. Specifically, modules with a blastz score over 2600 between medaka and human and with a percentage identity over 60% were ordered in descending order of module scores. BlastZ module homology searches in human were restricted to the orthologous neighborhood of each module, defined as follows. Each medaka module was first associated to the closest medaka gene. The human ortholog H, and the human genes flanking H on the left and the right were identified. The orthologous neigborhood was then defined as the region between these two flanking genes. From this list, we selected two datasets : [1] the top 10 scoring modules

and [2] 10 modules distributed at regular intervals in the top 200 scoring modules (module at position 20, 40, 60, 81, 100, 120, 140, 159, 180, 200)

4.6.2 In-situ enrichment analysis

Each predicted CRM is associated to the closest gene independently of the distance. We took advantage of the large collection of genes with zebrafish in-situ annotation available from the Zn in-situ database [Sprague *et al.*, 2006]. In order to transfer zebrafish gene in-situ annotation to medaka, we transferred zebrafish in-situ annotation to the medaka orthologs available from the BioMart utility [Flicek *et al.*, 2008, Kasprzyk *et al.*, 2004]. If more than one ortholog was available for a given zebrafish gene the ortholog with the highest identity was conserved. For each tissue (and its subparts) each stage specific to this tissue, we retrieved all of the genes expressed in this tissue and all CRMs are linked to the closest gene. We then calculate the significance of the over-representation of genes attached to vertebrate conserved CRMs for a specific tissue with the over-representation of genes attached to non-vertebrate conserved CRMs for the same tissue. The significance of this over-representation was calculated with a one-sided Fisher test. All tissue and stage annotations follow the OBO ontology.

4.6.3 Location analysis

For each CRM, the distance from the annotated TSS of the nearest protein coding gene (as defined in Ensembl version 53) was retrieved and categorized into either less than 1kb, 1 to 10 kb, 10 to 100 kb or more than 100 kb distances. The CRMs are also assessed if they are localized in annotated exons, introns of protein coding genes (as defined in Ensembl version 53). One hundred randomisations

consisting of the same number of random locations (of the same size distribution) in the medaka genome as the number of CRMs in the real dataset has been produced. The same location analysis is then performed on these random datasets and the significance is calculated from these randomizations.

4.6.4 Molecular cloning

The indentified CRMs were PCR amplified (using LA-Taq polymerase, Takara Bio Inc.) from genomic medaka DNA and flanking HindIII restriction sites were introduced. After restriction digestion, the fragments were cloned into a pBlueScript-based transgenesis vector containing two recognition sites for the meganuclease ISce-I19 flanking a multiple cloning site followed by the core promoter Hsp70::GFP14 and an SV40 polyadenylation signal (clone available upon request). For testing the CRMs acting in concert with their endogenous promoter 1 kb upstream of the TSS of otp1 (chromosome:MEDAKA1:9:6162143:6163143) and lmo1 (chromosome:MEDAKA1:3:18015929:18016929, chromosome:MEDAKA1:3:18028785:18029785) respectively were PCR amplified (using LA-Taq polymerase, Takara Bio Inc.) and flanking Nco1 and HindIII restriction sites introduced (primer sequences SUP. LINK). The fragments were subcloned into the construct described above instead of the Hsp70 core promoter together with their corresponding CRM. As the TSS of lmo1 was not clearly annotated two different genomic loci were chosen where only the second one (chromosome:MEDAKA1:3:18028785:18029785) gave expression. All cloned constructs were verified by DNA sequencing.

4.6.5 Medaka injection and screening

Injections were done as described [Rembold *et al.*, 2006]. DNA was purified using the Maxiprep Kit (Qiagen) and injected at a concentration of 15 ng/l. A Leica fluorescent microscope (Leica MZFLIII) was used to examine GFP expression in live embryos. Injected embryos were analysed at different stages to determine the spatio-temporal pattern of GFP expression. The hsp70 core promoter being activated by temperature changes, the embryos were kept and examined at room temperature. Developmental stages were determined by morphological features as described by Iwamatsu [Iwamatsu, 2004].

4.6.6 Whole mount in-situ hybridization

For analysis of scamp1, fign(1 of 2), atg4c, gon3_oryla and kcnh7 expression patterns, fragments were PCR amplified from medaka cDNA (using Taq-Polymerase) and subcloned utilizing the TOPO TA Cloning Kit (Invitrogen). After verification by sequencing Digoxigenin incorporated antisense-RNA probes were generated by *in vitro* transcription with Sp6 or T7 RNA Polymerase (NEB). Probe preparation and whole mount in-situ hybridization were performed as described previously [Loosli *et al.*, 2001]. For the remaining genes analyzed, we could find at least one clone matching part of the transcript sequence in our in-house library (in pCMV-Sport6.1). In these cases, probes were generated by *in vitro* transcription with Sp6 or T7 RNA Polymerase directly from these clones.

4.6.7 Medaka annotation

The medaka nervous system ontology is derived from the Medaka fish anatomy and development OBO ontology (medaka_ontology.obo) and all the

descendent terms of nervous system at various stages were extracted. A total of 32 different terms were found and used for the controlled vocabulary annotation. Reporter gene expressions were found in 20 (62%) of these anatomical terms.

4.7 Acknowledgments

EM is supported by funding from Genome Quebec/Canada. TA and LE are supported by the FP7 CISSTEM grant. This work was also supported in part by the EMBO Short-Term Fellowships program (EM).

4.8 Supplementary figures and tables

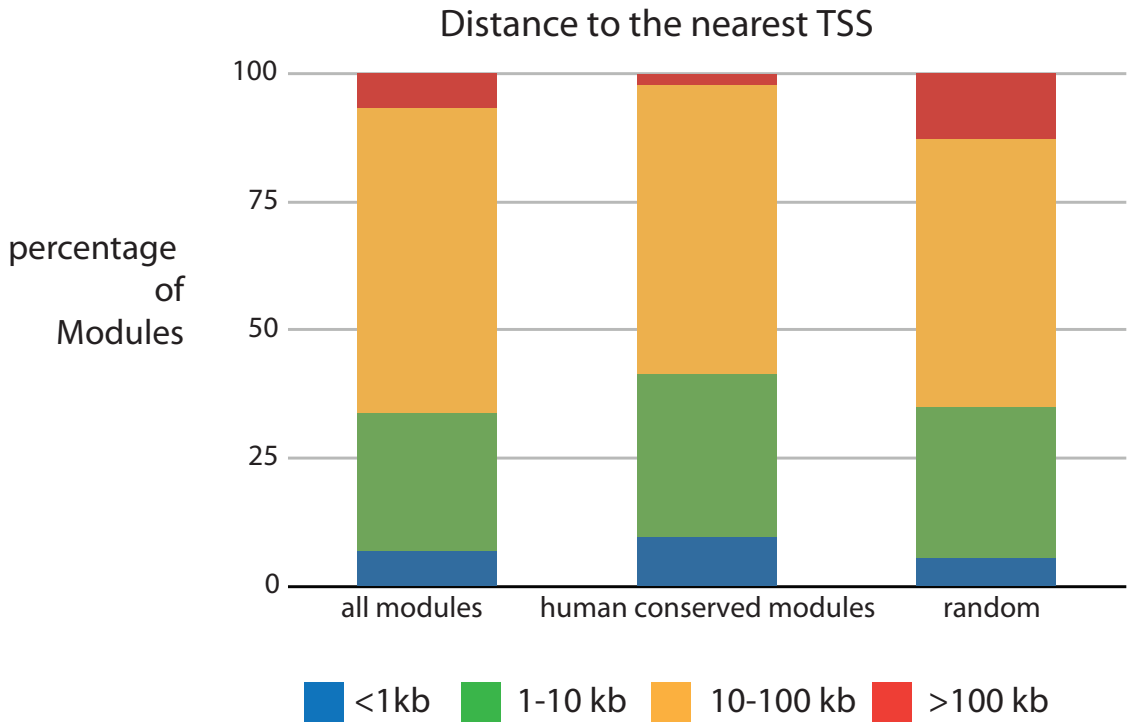

Figure 4–5: **Locations of the CRMs relative to the distance to the nearest annotated TSS.**

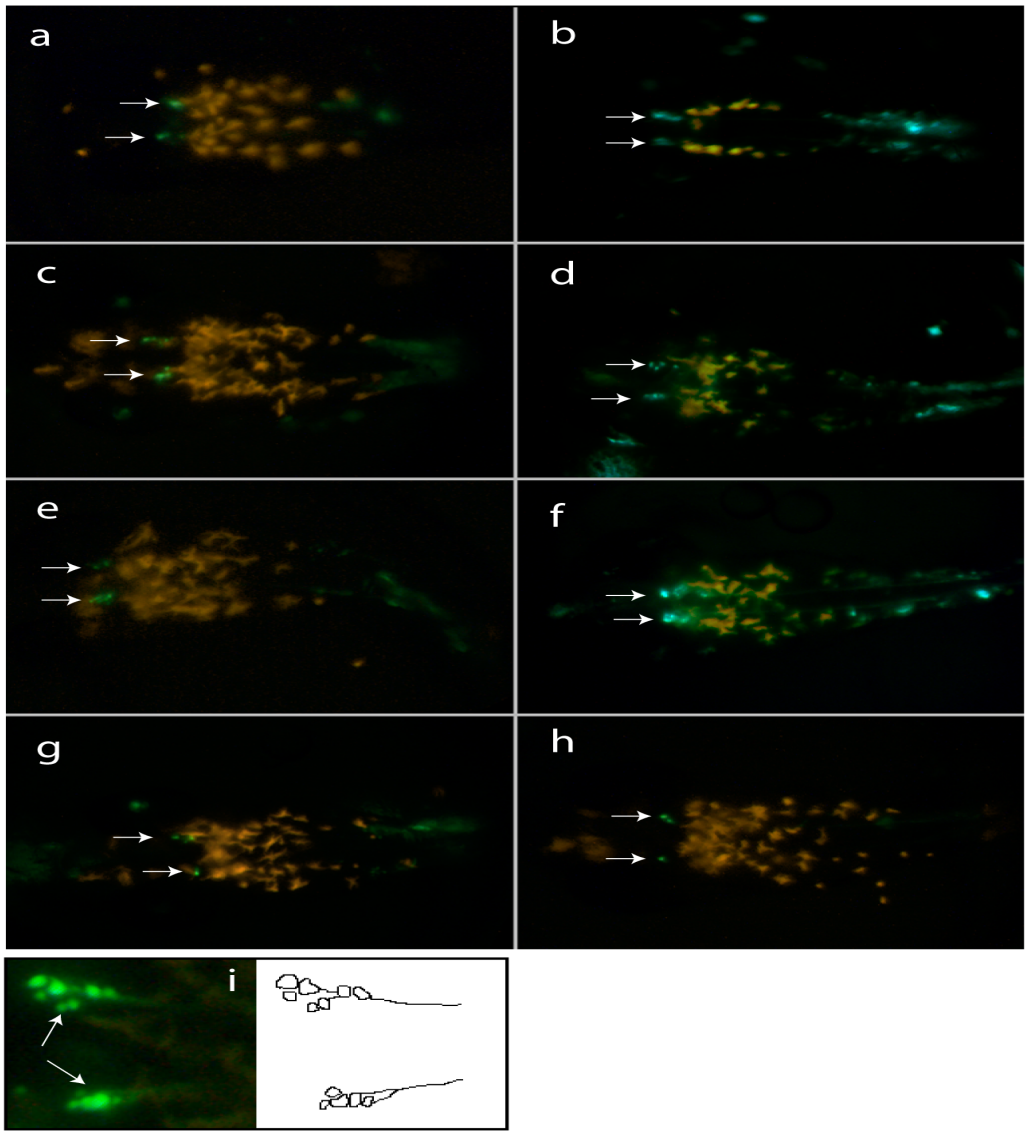

Figure 4–6: **detection of a restricted domain of expression in injected embryos.** (a-h) Different individual embryo stage 26 injected with the MEDMOD062451 construct. White arrows are indicating the GFP positive cells (diencephalon). i : Focus on the two populations of GFP positive cells in the diencephalon in the stable MEDMOD062451 line (stage 26). Even in stable lines, the number of labeled cells at this stage is limited. Despite the restricted domain of expression, the apropriate cells are labelled in most of the injected fish.

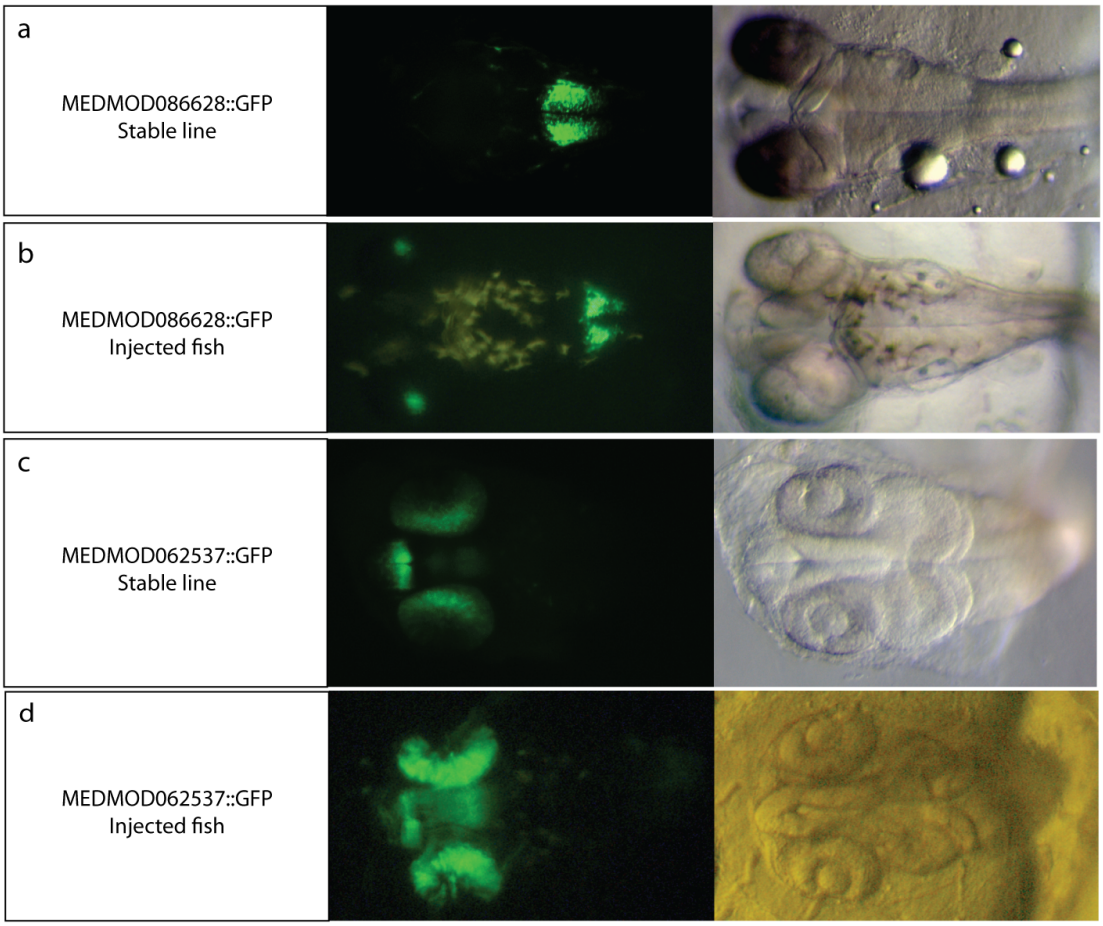

Figure 4–7: **Reporter gene expression in transient versus stable lines.** (a) MEDMOD086628-hsp70::GFP construct in stable line and (b) injected embryos. (c) MEDMOD62537-hsp70::GFP construct in stable line and injected embryos (d). The expression pattern of the reporter gene is similar in injected embryos and in stable lines indicative of a fast and efficient integration of the construct in the host genome.

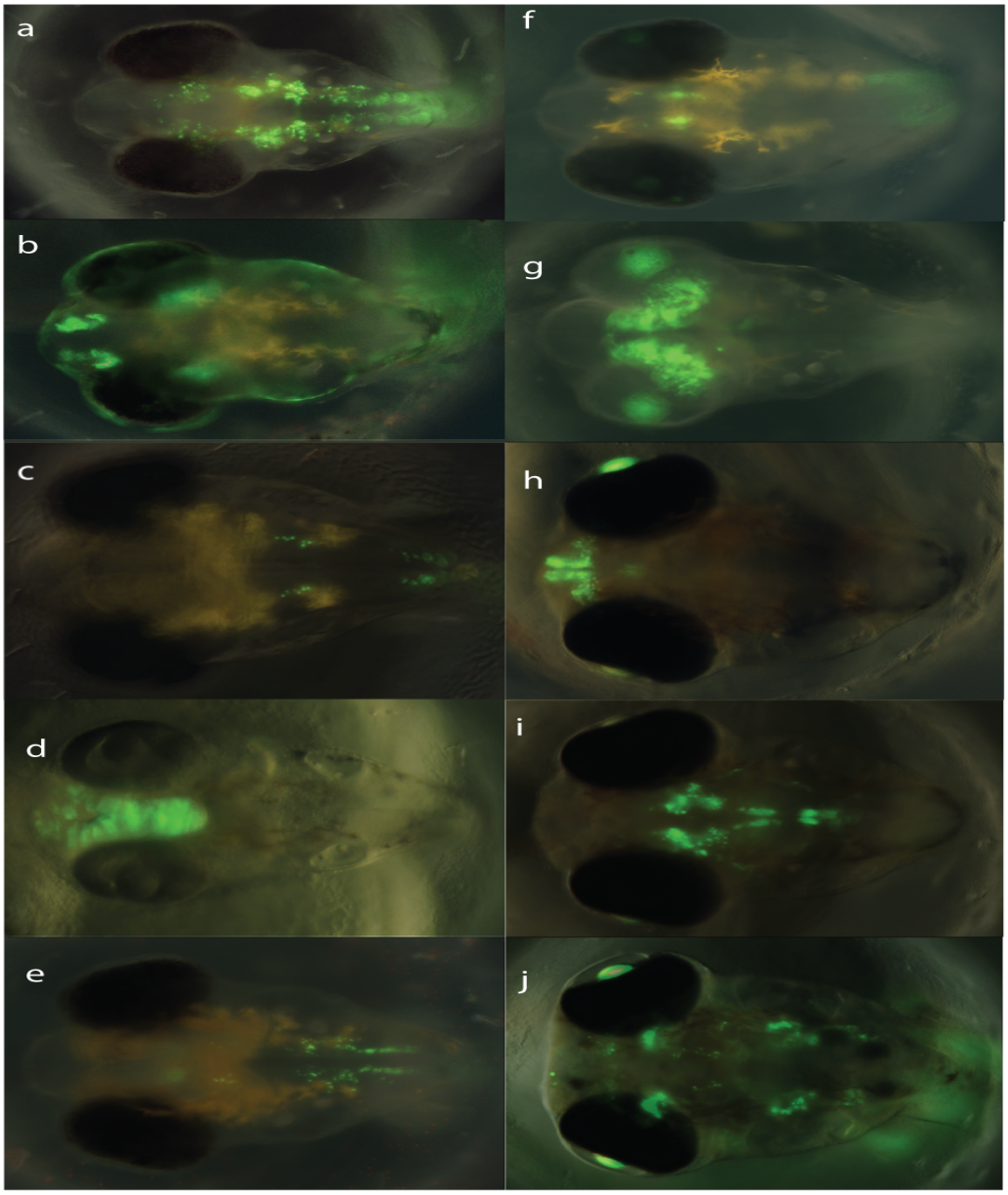

Figure 4–8: **10 constructs tested at different score levels.** (a) MEDMOD021445 (b) MEDMOD092210 (c) MEDMOD062490 (d) MEDMOD057815 (e) MEDMOD021442 (f) MEDMOD093196 (g) MEDMOD062408 (h) MEDMOD047799 (i) MEDMOD083481 (j) MEDMOD062206.

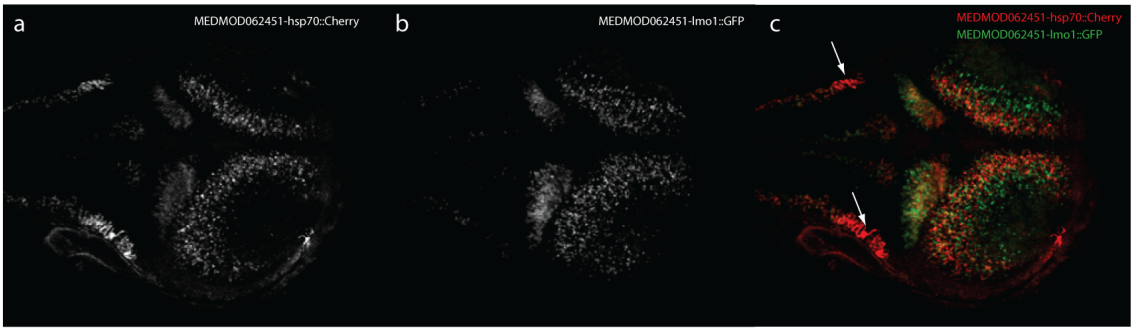

Figure 4–9: **Effect of the minimal promoter on reporter gene expression.** MEDMOD062451-hsp70::CHERRY, MEDMOD062451-lmo1::GFP double transgenic line (stage 33) (a) GFP expression pattern (b) CHERRY expresssion pattern (c) Merged image. The domains of expression of GFP are similar to the domains of expression of CHERRY. CHERRY expression extend further in the rhombic lips compare to GFP (arrows) suggesting a minimal effect of the promoter hsp70 on reporter gene expression pattern.

Table 4–1: **Enrichment of genes express in neuronal tissues around vertebrate conserved CRMs.** For each selected developmental tissue, and stage percentage of genes expressed in the given tissue and percentage of the rest of genes linked to non-vertebrate conserved CMRs at the same conditions. The statistical significance is calculated with a one-sided fisher test.

Stage	Tissue	p-value	% of annotated modules in conserved set	% of annotated modules in background
Gastrula	neural plate	0.0001005	27.77777778	15.15353805
Gastrula	whole organism	0.5986	75.69444444	76.18205921
Segmentation	hindbrain	6.83E-06	50	33.7158755
Segmentation	brain	4.86E-05	68.33333333	53.53851504
Segmentation	neural rod	0.000404	28.49162011	17.86984353
Segmentation	neural tube	0.000684	48.33333333	36.20239958
Segmentation	midbrain	0.001537	46.11111111	34.9504434
Segmentation	central nervous system	0.001834	68.08510638	57.30472103
Segmentation	neural keel	0.004318	41.37931034	31.48364486
Segmentation	presumptive diencephalon	0.00478	12.64367816	6.83411215
Segmentation	forebrain	0.007109	52.22222222	42.70561641
Segmentation	telencephalon	0.008847	33.33333333	25.05651191
Segmentation	anterior neural keel	0.03414	29.93630573	23.22702476
Segmentation	alar plate midbrain	0.0391	12.77777778	8.589810468
Segmentation	neural plate	0.05642	35.29411765	29.29757855
Segmentation	anterior neural rod	0.07059	12.35294118	8.712793219
Segmentation	pharyngeal arch	0.07768	13.33333333	9.772213528
Segmentation	diencephalon	0.08273	39.44444444	34.15058251
Segmentation	midbrain neural keel	0.09434	21.01910828	16.65967268
Segmentation	immature eye	0.1248	26.73796791	22.85075653
Segmentation	optic vesicle	0.1324	17.64705882	14.4383184
Segmentation	ectoderm	0.1501	18.08510638	15.05579399
Segmentation	midbrain neural tube	0.1839	21.66666667	18.72717788
Segmentation	hindbrain neural plate	0.1967	13.52941176	11.14668801
Segmentation	posterior neural plate	0.1967	13.52941176	11.14668801
Segmentation	anterior neural tube	0.2208	21.66666667	19.10972005
Segmentation	nervous system	0.3029	13.82978723	12.34334764
Segmentation	somite	0.393	22.34042553	21.28755365
Segmentation	head	0.4163	26.06382979	25.15021459
Segmentation	cranium	0.4168	16.66666667	15.85811163
Segmentation	neuron	0.4343	11.70212766	11.10729614
Segmentation	notochord	0.6163	15.95744681	16.54935622
Segmentation	basal plate midbrain	0.6271	12.22222222	12.81516258
Segmentation	trunk	0.8888	26.06382979	29.9055794
Segmentation	mesoderm	0.9154	26.06382979	30.43776824
Segmentation	whole organism	0.9949	33.5106383	42.54077253
Larval	midbrain	6.22E-05	50	29.13957935
Larval	hindbrain	0.000818	39.28571429	23.09751434
Larval	brain	0.00961	69.04761905	55.71701721
Larval	central nervous system	0.01196	71.42857143	58.6998088
Larval	forebrain	0.01361	45.23809524	32.88718929
Larval	head	0.07438	55.95238095	47.34225621
Larval	diencephalon	0.07519	27.38095238	20.22944551
Larval	eye	0.0985	46.42857143	38.81453155
Larval	visual system	0.101	46.42857143	38.89101338
Larval	retina	0.1281	42.85714286	36.17590822
Larval	optic tectum	0.1495	26.19047619	20.87954111

Table 4–2: **Injection success rate.** Alive column corresponds to the number of injected embryos which passed gastrulation. Expression corresponds to the number of embryos with expression pattern in the eyes (successful injection) and specific expression corresponds to the number of embryos with reproducible expression pattern excluding eye specific pattern.

Construct	Injected Eggs	Alive	Expression	Specific Expression	% of Expression	Percentage of Specific Expression
HSP	259	189	101	101	53.44	100.00
MEDMOD021953-HSP	178	84	32	20	38.10	62.50
MEDMOD062451-HSP	294	233	67	48	28.76	71.64
MEDMOD074008-HSP	157	111	53	35	47.75	66.04
MEDMOD021885-HSP	297	249	84	27	33.73	32.14
MEDMOD070042-HSP	166	139	65	36	46.76	55.38
MEDMOD046007-HSP	167	119	43	20	36.13	46.51
MEDMOD046561-HSP	74	69	32	0	46.38	0.00
MEDMOD045693-HSP	180	88	19	12	21.59	63.16
MEDMOD086628-HSP	97	76	35	26	46.05	74.29
MEDMOD062537-HSP	205	151	56	32	37.09	57.14

Transition to next Chapter

In this Chapter, we described a novel method to characterize *in vivo cis*-regulatory regions in medaka fish. This concludes the set of three manuscripts that compose the core of this thesis. In the next Chapter, "Genome plasticity and gene regulation", we discuss, with the support of the work previously described in this thesis and the literature, how long-range regulation may have influenced the shaping of the human throughout evolution.

CHAPTER 5
Genome plasticity and gene regulation

Evolutionary rearrangements occur at specific locations on the human genome. Although the opposite view [Nadeau & Taylor, 1984] remained prominent for decades, it is now widely accepted that breakpoints are fixed non-randomly throughout evolution [Pevzner & Tesler, 2003, Murphy *et al.*, 2005, Peng *et al.*, 2006]. However, the reasons remain poorly understood.

Firstly, we hypothesize that the primary force leading to different breakpoint susceptibility is the functional pressure imposed by long-range regulation on the genome structure. The expression of certain categories of genes such as developmental genes necessitates precise spatio-temporal regulation. This accuracy results from intricate functional links between *cis*-regulatory regions and target genes. Therefore, functional links cannot be broken without strong consequences on fitness. We describe in this chapter an intricate interplay between evolutionary pressures that restrict rearrangements to certain regions, where they stimulate new repeats, that in their turn favor recombination events.

Secondly, the existence of regions with such divergent evolutionary scenarios may impact the incidence of diseases that are related to chromosomal rearrangements. We discuss evidences that link evolutionary breakpoints to disease rearrangements and focus on two types of disease related rearrangements: i/ cancer breakpoints, and ii/ position-effect related rearrangements. With the support

of these two examples, we suggest that regions of the human genome with high evolutionary plasticity are particularly susceptible to disease rearrangements.

To summarize, we show how evolutionary pressures that are generated by long-range regulation participated in the creation of evolutionary divergent regions with distinct susceptibility to rearrangements. We hypothesize that such contrasting evolutionary plasticity directs disease related rearrangements to certain regions of the genome.

5.1 An intricate evolutionary interplay

5.1.1 Long-range regulation plays a role in genome stability

In various regions of the human genome, synteny of coding and non-coding regions is conserved within vertebrates. Such synteny may be the consequence of long-range regulation. Properties of long-range regulators favor this hypothesis: i/ they can act over long distance [Lettice *et al.*, 2003], and ii/ one regulatory region may regulate many genes [Spitz *et al.*, 2003]. Physical disruption of the functional link between *cis*-regulatory regions and target genes may be too detrimental to the fitness of the individual and not fixed in evolution [Mackenzie *et al.*, 2004, Mongin *et al.*, 2009] (see also Chapter 2). This is particularly important in regions necessitating precise spatio-temporal regulation such as developmental regions. Therefore, some regions of the genome, of the early vertebrate ancestor, were already refractory to rearrangements, whereas in other regions such rearrangements would not be as detrimental and therefore more common.

Such role in genome stability is exemplified in the work presented in Chapter 4. The set of teleost *cis*-regulatory regions that we characterized *in vivo* to

be involved in nervous tissue development are all in synteny with the same target genes in human. The proper shaping of nervous tissues in development is fundamental to the survival of the organism and no rearrangements may occur in those regions without dramatic consequences.

5.1.2 Conservation of synteny favors the recruitment of new regulatory elements

Evolution does not follow a conscious design but is mainly the consequence of "evolutionary tinkering" and adaptation [Duboule & Wilkins, 1998, Kleinjan & van Heyningen, 2005]. Therefore, conservation of synteny could favor the exaptation of new *cis*-regulatory elements. New expression patterns may result from the recruitment or creation of new *cis*-regulatory regions, given that they are able to interact with the promoter of the target gene and do not affect the existing regulation and fitness [Kleinjan & van Heyningen, 2005]. Such acquisition is shown by Spitz *et* al [Spitz *et al.*, 2001] at the HoxD cluster. They show that *cis*-regulatory regions involved in metazoan axial patterning are located within the cluster, and that functions specifically acquired by vertebrates (in limb and genitalia) are regulated by regions located outside of the cluster. They explain such acquisition either by the modification of trans-factors to bind those new *cis*-element, or by sequence changes of the newly acquired *cis*-regulatory regions. It is likely that the evolutionary stability of certain regions favors the co-evolution of non-coding regions and genes, and therefore the exaptation of new *cis*-regulatory elements by those genes in order to acquire new expression profiles and consequently new functions. It is also becoming clear that transposon elements may also be exaptated as new *cis*-regulatory regions [Bejerano *et al.*, 2006, Bourque, 2009].

Lowe *et al.* showed that non-coding conserved regions originating from mobile elements and under purifying selection are in the vicinity of developmental genes. Such recruitments augments the interdependency between genes and *cis*-regulatory regions, thus the cost to break these regions.

5.1.3 Local features associated with evolutionary breakpoints reflect evolutionary plasticity

Non-allelic homologous recombination (NAHR) and non-homologous end-joining (NHEJ) are among the two main mechanisms proposed to explain evolutionary rearrangements (see Chapter 1 for more details). Although the exact mechanisms remain under study, both of these processes seem to be linked to replication stress and repeats. The probability for NAHR to occur is closely correlated with the presence of sequences of high identity that may act as substrates to facilitate the recombination [Gu *et al.*, 2008]. Although NHEJ does not require repeat sequences to mediate the recombination, they may be stimulated by repeats such as low copy repeats (LCRs) [Stankiewicz *et al.*, 2003, Shaw & Lupski, 2004].

Consequently, it is not surprising to observe that evolutionary breakpoints predominantly lie in regions enriched for repeat features. For example, different types of repeats such as segmental duplications (SDs), (LINEs), short interspersed nuclear elements (SINEs), long terminal repeats (LTRs), α-satellites and $(AT)_n$ repeats overlap primate specific evolutionary breakpoints [Kehrer-Sawatzki & Cooper, 2008]. We also showed in Chapter 2 that regions classified as susceptible to evolutionary breakpoints are enriched for copy number variations (CNVs) compared to breakpoint-refractory regions. Causal effects have been proposed for certain types of repeats such as LINE or *Alu* elements, which

may be responsible for chromosomal inversions between human and chimpanzee [Lee *et al.*, 2008].

However, these repeated regions may not only be the cause but also the consequence of these rearrangements. For example, SDs are associated with 98% of primate-specific breakpoints [Murphy *et al.*, 2005] and from 25 to 53% of human/mouse breakpoints [Bailey *et al.*, 2004, Armengol *et al.*, 2003] (reviewed in [Kehrer-Sawatzki & Cooper, 2008]). However, Bailey *et al.* correlated mouse-lineage specific rearrangements with 30-35 million years old primate specific SDs. This implies that SDs are probably not the cause for primate evolutionary rearrangements [Bailey *et al.*, 2004], but the consequence of genome plasticity. Interestingly, the burst of *Alu* mediated retransposition about 35 million years ago has been linked to the primate expansion of SDs. Bailey *et al.* proposed that the high sequence identity of *Alu* sequences may favor NAHR and consequently the spread of SDs [Bailey *et al.*, 2003].

We hypothesize that regions of high genomic plasticity originated from genomic regions with low regulatory complexity. Lower functional pressure favored the creation of repetitive elements such as transposon associated repeats, which in turn act as substrate for subsequent rearrangements such as NAHR. Therefore throughout evolution, these regions became highly plastic regions exemplified by the burst of SDs in primate evolution. Some genes seem to benefit from such genomic plasticity. Regions that are rearranged in evolution are enriched for genes involved in adaptive processes such as immune response and, response to external stimuli (see Chapter 3, [Larkin *et al.*, 2009]). Such classes of genes need to be

extremely adaptive the to external environment and it is not surprising that genes carrying such functions lie in regions that are prone to be rearranged.

5.2 Evolutionary and disease breakpoints

5.2.1 Evolutionary breakpoints and cancer rearrangements

Are evolutionary breakpoints and disease rearrangements such as cancer rearrangements driven by the same mechanisms and localized within the same regions of instability? Different studies have shown an overlap between evolutionary breakpoints and cancer rearrangements. Among such studies, Murphy *et al.* [Murphy *et al.*, 2005] noticed that highly prevalent cancer rearrangements are enriched within mammalian evolutionary breakpoints compared to low prevalence rearrangements. Studies focussing on chromosome 3 showed a correlation between tumor break-prone regions and evolutionary rearrangements [Kost-Alimova *et al.*, 2003, Ruiz-Herrera & Robinson, 2008, Darai-Ramqvist *et al.*, 2008].

Evolutionary rearrangements and cancer breakpoints are driven by different forces, since evolutionary rearrangements need to be neutral or to confer a fitness advantage to the individual whereas cancer breakpoints occur in somatic cells and may confer a growth advantage to the cell (by disrupting a tumor suppressor gene for example). However, such overlaps raise the possibility that both types of rearrangements have affinity to similar regions (but not especially to the same exact location). Moreover, NAHR and NHEJ occur both in somatic and germline cells [Gu *et al.*, 2008] and may therefore be responsible for both cancer and evolutionary rearrangements.

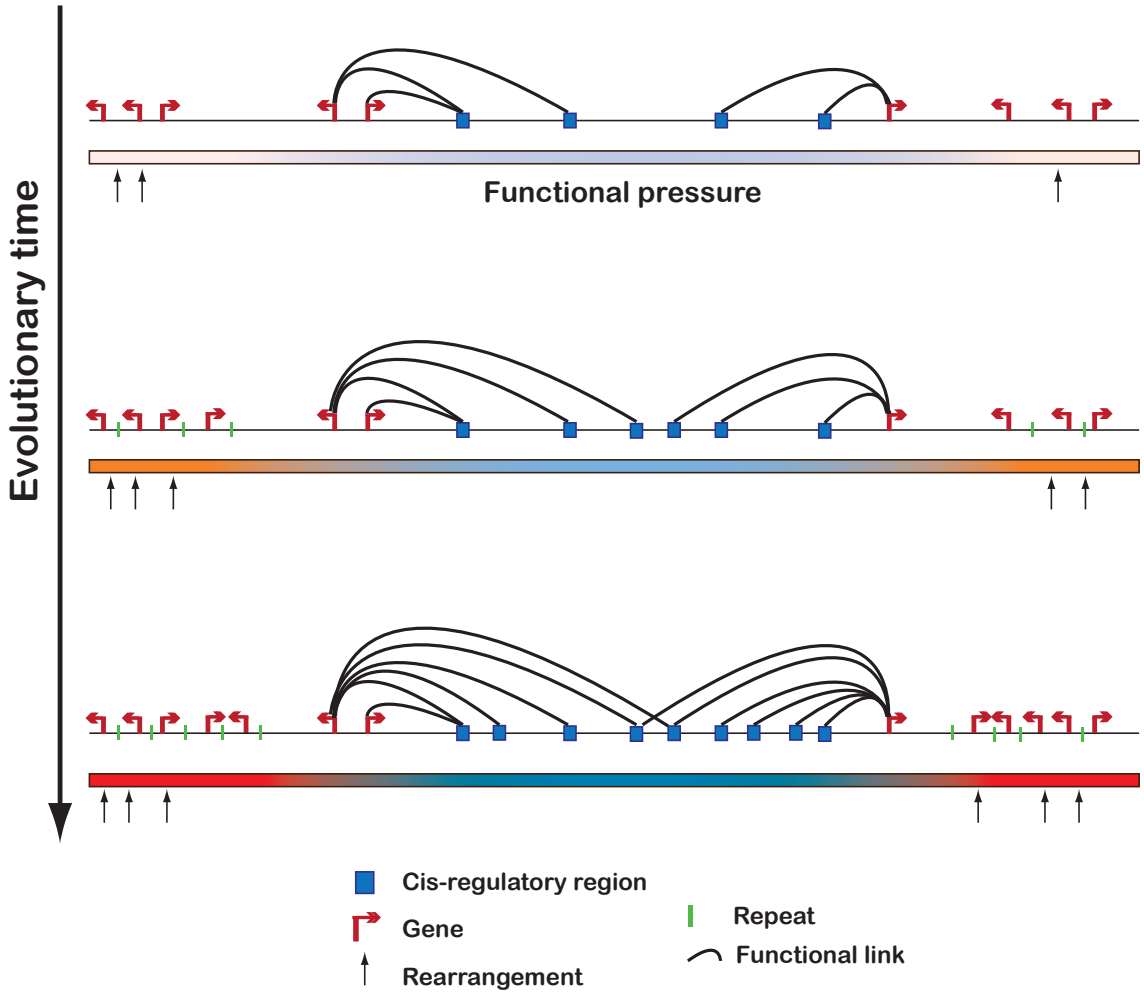

Figure 5–1: **Schematic representation of functional pressure variations on the genome through evolution.** Three different representation of a fictive genomic region are presented: i/ before teleost divergence, ii/ before mammalian divergence, and iii/ before hominoid divergence. Functional pressure on different parts of the region is illustrated with a color gradient, where blue represents regions where functional pressure prevents rearrangements and red where rearrangements are tolerated. Darker blue or red color symbolizes respectively stronger functional pressure and more important plasticity.

In our model, we propose that genomic features linked to plastic regions may favor the occurrence of cancer-related rearrangements. If this model is correct, highly prevalent cancer rearrangements should preferentially occur in these regions. Inspired by an analysis briefly mentioned in [Murphy *et al.*, 2005], we associated the predictions presented in Chapter 3 (that to a certain extent reflect the evolutionary pressure on each gene with regards to regulation) with cancer breakpoint prevalence. We notice that cancer breakpoints observed in a small number of patients (under 5 reported cases) are preferentially linked to genes with many predicted functional regulatory links with *cis*-regulatory regions. However, cancer breakpoints observed in many patients (over 77 reported cases) mainly correspond to regions where genes have few functional regulatory links. Although this analysis was limited to a small number of disease associated breakpoints and probably conveys a simplified description of such complex mechanisms. These results tend towards our model since it shows that cancer breakpoints preferentially occur in the vicinity of genes that are not under complex regulation. Interestingly, no cancer breakpoints are observed within the three longest syntenic regions described by [Murphy *et al.*, 2005].

5.2.2 Disease-related position-effect rearrangements

Disease-related position-effect rearrangements are rearrangements that are located outside of coding loci and result in the mis-regulation of a gene. Most reported position-effect diseases involve genes that have well conserved biological processes such as developmental genes [Kleinjan & van Heyningen, 2005] and

have strong phenotypical effects. See Table 5–1 for examples of disease causing mutations.

Table 5–1: **Example of disease-causing mutations affecting long-range regulators.** Some examples from this table are directly taken from [Kleinjan & van Heyningen, 2005].

Disease name	Gene name	Mechanism	Reference
Aniridia	PAX6	Translocation	[Kleinjan *et al.*, 2001]
X-linked deafness	POU homeodomain	Chromosomal deletion	[de Kok *et al.*, 1996]
Campomelic dysplasia	SOX9	Translocation	[Bagheri-Fam *et al.*, 2001]
BPES	FOXL2	Translocation	[Crisponi *et al.*, 2004]
Acute myeloid leukemia	PU.1	SNP	[Steidl *et al.*, 2007]
Blood pressure	Renine	SNP	[Vangjeli *et al.*, 2010]

Our model indicates that developmental regions are particularly refractory to rearrangements as a consequence of the strong functional pressures at work to maintain genes and *cis*-regulatory regions in synteny. However, even though the local genomic context does not favor rearrangement in developmental regions, the possibility to observe rearrangements is not excluded but should be seldom. Therefore, is not surprising to observe that these diseases are almost entirely classified as rare diseases with strong phenotypical effects. For example, aniridia, Saethre-Chotzen syndrome or Rieger syndrome have a prevalence of about 1-9 in 100 000 (source: http://www.orpha.net).

5.2.3 Evolutionary breakpoint susceptibility scores may be useful to target rearrangements of interest

It is not obvious to know which rearrangements observed in patients are responsible for the disease under investigation. In chapter 2 we scored the human genome based on its tolerance to evolutionary breakpoints. These predictions reflect the effect that a rearrangement will have at a given genomic location on

the phenotype. Regions with high scores are regions that are prone to breakpoints and therefore regions where rearrangements may have limited phenotypical consequences. In contrast, regions with low scores correspond to regions that are refractory to rearrangements and therefore where rearrangements may have serious deleterious effects and be responsible for disease phenotypes.

We therefore propose that our predictions can be particularly useful to determine which chromosomal aberration in a patient may be responsible for the observed phenotype. Chromosomal aberration located in regions with low scores may be more likely to be involved in genetic diseases such as diseases related to development. As shown in chapter 2, regions with low scores correspond to regions that are enriched for genes involved in transcriptional and developmental processes. In contrast, chromosomal aberration located in high scoring regions are more likely to be silent. However, as described in Chapter 2, rearrangements within gene loci are really rare and gene loci would be considered as refractory to rearrangement even if located within high scoring regions. Rearrangements observed within gene loci of regions with high scores are more likely to disrupt genes that are involved in housekeeping function and may be more likely to lead to diseases such as cancers.

To conclude, we propose an evolutionary phenomenon where long-range regulation plays a major role in shaping the genome. On one hand, functional pressures imposed by long-range regulation generated the development of syntenic conserved regions. On the other hand, other regions developed local conditions favorable to the emergence of rearrangements. We suggest that such dichotomy

also affects the prevalence of disease-related rearrangements. Although this model may appear simplistic given the complex biological mechanisms put forward, we believe that it provides orientations to further research and will hopefully lead to a better understanding of the mechanisms responsible for rearrangement-related diseases.

CHAPTER 6
Conclusion

6.1 Summary and contribution to original knowledge

6.1.1 Long-range regulation is a major force in maintaining genome integrity

Since the landmark paper from Pevzner *et al.* [Pevzner & Tesler, 2003], it is widely accepted that evolutionary breakpoints occur in specific regions of the genome referred as "fragile regions". Although these regions have been associated with various genomic features such as repeats and increased gene density [Murphy *et al.*, 2005, Kehrer-Sawatzki & Cooper, 2008], the evolutionary forces that favor regional specificity are poorly understood. Long-range regulation was both proposed as a force responsible to maintain the synteny of some regions (such as the Hox cluster [Lee *et al.*, 2006] or the Shh locus [Goode *et al.*, 2005]) and as a more general mechanism [Mackenzie *et al.*, 2004, Kikuta *et al.*, 2007].

In **Chapter 2**, we took a purely computational approach and developed a new method to predict genomic susceptibility to evolutionary rearrangements. With this method we were able to: i/ define which genomic features favor breakpoints, and ii/ what are the specificities of regions susceptible and refractory to rearrangements. Our method is based on the training of a machine learning method on human-lineage specific breakpoints taken in their ancestral context with different genomic features. Our predictor performed well; using cross-validation,

we show that 75% of breakpoints are predicted in about 27% of inter-marker regions (see Figure 2–4). From these predictions, we defined two types of regions: i/ 'breakpoint-refractory regions' where rearrangements would be too deleterious to be fixed in evolution, and ii/ 'breakpoint-susceptible regions', where rearrangements are expected to be neutral.

Firstly, we showed that only a small fraction of the genome is susceptible to breakpoints. Secondly, we uncovered different functional characteristics between susceptible and refractory regions. On one hand, refractory regions are strongly enriched for non-coding conserved elements, genes involved in transcription and development and for tissue-specific genes (see Table 2–3). On the other hand, susceptible regions are enriched for genomic features related to chromosome instability such as CNVs and fragile sites (see Table 2–6). We claim that this functional dichotomy between breakpoint-susceptible and breakpoint-refractory regions reflects in part the importance of long-range regulation as a main mechanism to prevent rearrangements from being fixed in parts of the genome.

After the submission of **Chapter 2**, Larkin *et al.* published a paper, based on the analysis of rearrangements between 10 amniote genomes [Larkin *et al.*, 2009] using an approach similar to Murphy *et al.* [Murphy *et al.*, 2005]. They, similarly to us, compared regions of synteny to regions of evolutionary breakpoints and also provided original genome-wide comparison between these two types of regions. However, the advantages of our method contrasting with such analysis are that: i/ our model can be improved by adding features, and ii/ that our analysis is not limited to regions where breakpoints are detected. We started our analysis

with three types of elements (non-coding conserved regions, genes and NCE-genes associations), however additional genomic features that may relate to breakpoints (such as repeats) can be added to the predictor.

6.1.2 Evolutionary breakpoints as a tool to map regulatory domains

Long-range regulatory regions can regulate genes over large distance [Lettice *et al.*, 2003] and have the ability to control several genes [Spitz *et al.*, 2003]. Consequently, the prediction of target genes for *cis*-regulatory regions is a difficult task. Such a map would meet various interests and allow researchers i/ to make an educated choice at the time of prioritizing regions to be tested *in vivo* (in the context of position-effect related disease for example), ii/ to bring additional signals to large scale analysis such as the identification of regulatory SNPs.

In **Chapter 2** we showed that evolutionary breakpoints would preferentially occur in regions where no functional regulatory association is broken. Based on these observations, we developed a new method to predict putative functional interactions between *cis*-regulatory regions and genes. Our method is based on the assessment of the physical association of candidate pairs in human and 16 vertebrate genomes and scored with an EM-like algorithm. This method presented in **Chapter 3** provides association scores for 1,406,084 candidate associations.

These predictions will be submitted with the paper and therefore available to scientists that need to make educated guess prior to *in vivo* testing or add information to large scale analysis. This work is original since it does not simply reconstruct blocks of synteny to delimitate gene regulatory boundaries but assesses a score for each possible pair. Contrasting with other approaches such

as [Sun *et al.*, 2008], our method provides target gene predictions for NCEs at different levels of conservation that makes our predictions the most comprehensive set of NCE-target genes predictions published so far.

6.1.3 *In vivo* testing of position mutation candidates

Great efforts have been made to develop more specific and sensitive computational methods to predict putative *cis*-regulatory regions. However laboratory techniques are not adapted to keep up with the increasing number of regions to be tested. Some large scale methods such as ChIP-CHIP [Ren *et al.*, 2000] exist but are limited to specific transcription factors. In the context of position-effect related diseases, such as aniridia [Kleinjan *et al.*, 2001], Pierre Robin sequence [Benko *et al.*, 2009] or Campomelic displasia [Pop *et al.*, 2004] where affected genes are often key developmental genes, robust methods to characterize *in vivo* regulatory regions putatively involved in the disease is crucial.

The main challenge when designing such new method is to have a high confidence that the expression pattern observed corresponds to the actual endogenous expression specificity of the putative *cis*-regulatory region tested. The expression pattern can be altered in different manners: i/ The minimal promoter may produce additional expression patterns or distort the expression specificity of the regulatory regions under study. ii/ The reporter gene may be under the influence of other regulatory regions present at the point of insertion of the construct (position effect).

In **Chapter 4**, we propose a combined approach with *in silico* predictions and *in vivo* characterization of predicted *cis*-regulatory regions in medaka fish.

The method that we developed can be applied to medium scale testing (of the order of a few hundred of regions) and show the characteristics of a reliable tool. Among these characteristics, we report: i/ a high reproducibility, and ii/ a minimal promoter that mimic the endogenous promoter. We tested 20 CRMs at different levels of scores from a subset of the predictions that are conserved with human. Our prediction pipeline appears to be particularly effective since 95% of the 20 predicted regions we tested show expression specificity. We also report that 100% of the tested predictions that showed expression triggered nervous tissue expression, observation supported by a computational analysis of the 500 human conserved CRMs. This characteristics is particularly valuable to decipher the developing vertebrate nervous system as reported Chapter 4. Finally, we created transgenic lines for 10 of the constructs and obtained a 100% success rate. This characteristic is particularly important to precisely characterize expression pattern and if those transgenic fish are to be used as markers for genetic experiments.

To conclude, this method with such combined reliability and specificity, is a step forward in the *in vivo* characterization of putative regulatory regions and we believe that the tool we provide will become a reference for *cis*-regulatory regions testing in teleost fish. We also uncovered a strong correlation both computationally and *in vivo* between conservation and specificity of regulatory regions for nervous tissue development.

6.1.4 Thesis work in the context of position-effect related diseases

With this subsection, we aim at showing the use of the methods developed in this thesis in the context of the study of position-effect related disease. We

take here the example of blepharophimosis/ptosis/epicanthus inversus syndrome (BPES), a rare genetic disorder.

Misregulation or mutation of FOXL2 gene is responsible for BPES. Crisponi *et* al [Crisponi *et al.*, 2004] reported three translocations, 171 kb 5' upstream of FOXL2 transcription start site, responsible for BPES. They report some conserved regions whose functional link has been disrupted by the translocation (located over 171 kb upstream from the FOXL2 TSS) but do not propose any functional testing plan. In the context of such position mutation diseases - which are most of the time located in evolutionary conserved region [Kleinjan & van Heyningen, 2005] - the combined approach of our method to assess regulatory regions boundaries with the *in vivo* approach in teleost fish is particularly adapted. In this specific example, 42 conserved regions located between 171 kb and 1 Mb have from our prediction scores between -9 and 101 among which, 6 with an association score over 49. Within this subgroup, three regions are of particular interest located 281 569, 283 721, and 283 892 base pairs, 5' upstream from the TSS, with respective scores of 97, 101, and 101. Testing *in vivo* these 9 regions with our system in teleost is only a matter of weeks. If the functional link between FOXL2 and these putative regulatory regions cannot stricly been made using such an approach, comparison of *in situ* expression patterns from the gene in the developing embryo and expression patterns from the region tested is a good indication for functional linkage.

6.2 Summary of major contributions

We can summarize the most important contributions of this thesis as follows:

- A new insight on the influence of long-range regulation in shaping the human genome.
- A novel approach to predict regions susceptible to evolutionary rearrangements.
- An algorithm to assess the likelihood of functional association between *cis*-regulatory regions and genes.
- A novel biomolecular method to reliably characterize *cis*-regulatory regions in medaka fish.
- A prediction dataset of about 500 putative *cis*-regulatory regions highly enriched for enhancers with nervous tissue regulatory specificity.

6.3 Perspectives

The predictor we presented in **Chapter 2** was designed with a focus on long-range regulation. Therefore with a limited set of features related to regulation, we uncovered at the genome scale how long-range regulation is a major component in shaping the human genome. Our method can be adapted in various manners: i/ First, additional genomic features known to be related to evolutionary breakpoints (such as repeats) can be added to the model and would most probably increase the accuracy of the predictions. ii/ Second, other kinds of rearrangements could be investigated using the model. For example, the Mittelman database (http://cgap.nci.nih.gov/Chromosomes/Mitelman) contains 57,402 chromosome aberrations from patients. These aberrations (and their prevalence) are described in their cytogenetic context and are linked to different types of cancers. Our approach applied to theses somatic rearrangements would most probably lead to the

identification of new regions prone to cancer rearrangement, although resolution may be an issue.

In **Chapter 3**, we developed a novel method to predict functional interactions between candidate pairs composed of a gene and a non-coding conserved element. For a significant number of candidate associations there is not enough evolutionary evidence to make a decision and therefore additional genomes would be beneficial to produce a more comprehensive dataset. Our method is easily scalable to new genomes and such addition when genomes become available will particularly improve the predictions of pairs whose members are less conserved. The next step would be to confirm our predictions with experimental evidences. For example with the development of methods such as the 5C technology (see Chapter 1 for more details), large scale chromatin interaction datasets will most probably be available in the near future. Such datasets will be extremely useful to evaluate the extent of overlap between our predictions and physical chromatin interaction and therefore properly assess the biological significance of our predictions.

The method presented in **Chapter 4** can be applied to a whole range of applications. We showed from computational analysis and *in vivo* testing that the set of about 500 CRMs we provide will most probably trigger expression in various sub-regions of vertebrate nervous tissues. Comprehensive *in vivo* testing of those regions will therefore uncover new anatomical divisions of the nervous tissues and should be of great interest to developmental neurobiologists. We showed that it is straightforward to produce transgenic lines of medaka fish that carry the construct. Our method is therefore convenient to produce transgenic lines that express GFP

in specific nervous tissues. Such lines would be extremely valuable in the context of genetic experiments, for example to monitor the effect of a gene knockout on the development of specific nervous tissues. More generally, we believe that the method we develop will facilitate the characterization of *cis*-regulatory regions involved in vertebrate development.

References

[Agranat *et al.*, 2008] Agranat, L., Raitskin, O., Sperling, J., & Sperling, R. (2008). The editing enzyme adar1 and the mrna surveillance protein hupf1 interact in the cell nucleus. Proc Natl Acad Sci USA *105*(13), 5028–33.

[Aguilera & Gómez-González, 2008] Aguilera, A. & Gómez-González, B. (2008). Genome instability: a mechanistic view of its causes and consequences. Nat Rev Genet *9*(3), 204–17.

[Ahituv *et al.*, 2005] Ahituv, N., Prabhakar, S., Poulin, F., Rubin, E. M., & Couronne, O. (2005). Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. Hum Mol Genet *14*(20), 3057–63.

[Al-Shahrour *et al.*, 2008] Al-Shahrour, F., Carbonell, J., Minguez, P., Goetz, S., Conesa, A., Tárraga, J., Medina, I., Alloza, E., Montaner, D., & Dopazo, J. (2008). Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. Nucleic Acids Res *36*(Web Server issue), W341–6.

[Amacher, 1999] Amacher, S. L. (1999). Transcriptional regulation during zebrafish embryogenesis. Curr Opin Genet Dev *9*(5), 548–52.

[Amsterdam & Becker, 2005] Amsterdam, A. & Becker, T. S. (2005). Transgenes as screening tools to probe and manipulate the zebrafish genome. Dev Dyn *234*(2), 255–68.

[Antequera & Bird, 1993] Antequera, F. & Bird, A. (1993). Number of cpg islands and genes in human and mouse. Proc Natl Acad Sci USA *90*(24), 11995–9.

[Aparicio *et al.*, 2002] Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M., Roach, J., Oh, T., Ho, I., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S., Clark, M., Edwards, Y., Doggett, N., Zharkikh, A., Tavtigian, S., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., & Brenner, S. (2002). Whole-genome shotgun assembly and

analysis of the genome of fugu rubripes. Science *297*(5585), 1301–10. 1095-9203 (Electronic) Journal Article.

[Armengol *et al.*, 2003] Armengol, L., Pujana, M. A., Cheung, J., Scherer, S. W., & Estivill, X. (2003). Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. Hum Mol Genet *12*(17), 2201–8.

[Arnone & Davidson, 1997] Arnone, M. & Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. Development *124*(10), 1851–64. 0950-1991 (Print) Journal Article Review.

[Atchison, 1988] Atchison, M. L. (1988). Enhancers: mechanisms of action and cell specificity. Annu Rev Cell Biol *4*, 127–53.

[Bacolla *et al.*, 2004] Bacolla, A., Jaworski, A., Larson, J. E., Jakupciak, J. P., Chuzhanova, N., Abeysinghe, S. S., O'Connell, C. D., Cooper, D. N., & Wells, R. D. (2004). Breakpoints of gross deletions coincide with non-b dna conformations. Proc Natl Acad Sci USA *101*(39), 14162–7.

[Bagheri-Fam *et al.*, 2006] Bagheri-Fam, S., Barrionuevo, F., Dohrmann, U., Günther, T., Schüle, R., Kemler, R., Mallo, M., Kanzler, B., & Scherer, G. (2006). Long-range upstream and downstream enhancers control distinct subsets of the complex spatiotemporal sox9 expression pattern. Dev. Biol. *291*(2), 382–97.

[Bagheri-Fam *et al.*, 2001] Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G., & Pfeifer, D. (2001). Comparative genomics of the sox9 region in human and fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. Genomics *78*(1-2), 73–82.

[Bailey *et al.*, 2004] Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D., & Eichler, E. E. (2004). Hotspots of mammalian chromosomal evolution. Genome Biol. *5*(4), R23.

[Bailey *et al.*, 2003] Bailey, J. A., Liu, G., & Eichler, E. E. (2003). An alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet *73*(4), 823–34.

[Beaudoing & Gautheret, 2001] Beaudoing, E. & Gautheret, D. (2001). Identification of alternate polyadenylation sites and analysis of their tissue distribution using est data. Genome Res *11*(9), 1520–6.

[Becker & Lenhard, 2007] Becker, T. S. & Lenhard, B. (2007). The random versus fragile breakage models of chromosome evolution: a matter of resolution. Mol Genet Genomics *278*(5), 487–91.

[Bejerano *et al.*, 2004a] Bejerano, G., Haussler, D., & Blanchette, M. (2004a). Into the heart of darkness: large-scale clustering of human non-coding dna.. Bioinformatics *20 Suppl 1*, I40–I48.

[Bejerano *et al.*, 2006] Bejerano, G., Lowe, C., Ahituv, N., King, B., Siepel, A., Salama, S., Rubin, E., Kent, W., & Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature *441*(7089), 87–90. 1476-4687 (Electronic) Journal Article.

[Bejerano *et al.*, 2004b] Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W., Mattick, J., & Haussler, D. (2004b). Ultraconserved elements in the human genome. Science *304*(5675), 1321–5. 1095-9203 (Electronic) Journal Article.

[Benko *et al.*, 2009] Benko, S., Fantes, J. A., Amiel, J., Kleinjan, D.-J., Thomas, S., Ramsay, J., Jamshidi, N., Essafi, A., Heaney, S., Gordon, C. T., McBride, D., Golzio, C., Fisher, M., Perry, P., Abadie, V., Ayuso, C., Holder-Espinasse, M., Kilpatrick, N., Lees, M. M., Picard, A., Temple, I. K., Thomas, P., Vazquez, M.-P., Vekemans, M., Crollius, H. R., Hastie, N. D., Munnich, A., Etchevers, H. C., Pelet, A., Farlie, P. G., Fitzpatrick, D. R., & Lyonnet, S. (2009). Highly conserved non-coding elements on either side of sox9 associated with pierre robin sequence. Nat. Genet. *41*(3), 359–64.

[Bernstein *et al.*, 2007] Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The mammalian epigenome. Cell *128*(4), 669–81.

[Bird, 1995] Bird, A. P. (1995). Gene number, noise reduction and biological complexity. Trends Genet. *11*(3), 94–100.

[Blanchette *et al.*, 2006] Blanchette, M., Bataille, A., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D., Coulombe, B., & Robert, F. (2006). Genome-wide computational prediction

of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res *16*(5), 656–68. 1088-9051 (Print) Journal Article.

[Blanchette *et al.*, 2004] Blanchette, M., Kent, W., Riemer, C., Elnitski, L., Smit, A., Roskin, K., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E., Haussler, D., & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res *14*(4), 708–15. 1088-9051 (Print) Journal Article.

[Blanchette *et al.*, 1999] Blanchette, M., Kunisawa, T., & Sankoff, D. (1999). Gene order breakpoint evidence in animal mitochondrial phylogeny. J Mol Evol *49*(2), 193–203.

[Blechinger *et al.*, 2002] Blechinger, S. R., Evans, T. G., Tang, P. T., Kuwada, J. Y., Warren, J. T., & Krone, P. H. (2002). The heat-inducible zebrafish hsp70 gene is expressed during normal lens development under non-stress conditions. Mech Dev *112*(1-2), 213–5.

[Boffelli *et al.*, 2003] Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K., Ovcharenko, I., Pachter, L., & Rubin, E. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science *299*(5611), 1391–4. 1095-9203 (Electronic) Journal Article.

[Bondarenko *et al.*, 2003] Bondarenko, V. A., Liu, Y. V., Jiang, Y. I., & Studitsky, V. M. (2003). Communication over a large distance: enhancers and insulators. Biochem Cell Biol *81*(3), 241–51.

[Bourque, 2009] Bourque, G. (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. Curr Opin Genet Dev *19*(6), 607–12.

[Brudno *et al.*, 2003a] Brudno, M., Do, C., Cooper, G., Kim, M., Davydov, E., Green, E., Sidow, A., & Batzoglou, S. (2003a). Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. Genome Res *13*(4), 721–31. 1088-9051 Evaluation Studies Journal Article.

[Brudno *et al.*, 2003b] Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., & Batzoglou, S. (2003b). Glocal alignment: finding rearrangements during alignment. Bioinformatics *19 Suppl 1*, i54–62.

[Burke & Baniahmad, 2000] Burke, L. J. & Baniahmad, A. (2000). Co-repressors 2000. FASEB J *14*(13), 1876–88.

[Bushati & Cohen, 2007] Bushati, N. & Cohen, S. M. (2007). microrna functions. Annu Rev Cell Dev Biol *23*, 175–205.

[Butler & Kadonaga, 2001] Butler, J. E. & Kadonaga, J. T. (2001). Enhancer-promoter specificity mediated by dpe or tata core promoter motifs. Genes Dev *15*(19), 2515–9.

[Cai *et al.*, 2004] Cai, X., Hagedorn, C. H., & Cullen, B. R. (2004). Human micrornas are processed from capped, polyadenylated transcripts that can also function as mrnas. RNA *10*(12), 1957–66.

[Carninci *et al.*, 2005] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Gatta, G. D., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Babu, M. M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic,

V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., Consortium, F., Group, R. G. E. R., & Group), G. S. G. G. N. P. C. (2005). The transcriptional landscape of the mammalian genome. Science *309*(5740), 1559–63.

[Carninci *et al.*, 2006] Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., & Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution.. Nat. Genet. *38*(6), 626–35.

[Cawley *et al.*, 2004] Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., & Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. Cell *116*(4), 499–509.

[Chang *et al.*, 2006] Chang, L.-W., Nagarajan, R., Magee, J. A., Milbrandt, J., & Stormo, G. D. (2006). A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. Genome Res *16*(3), 405–13.

[Choo *et al.*, 2006] Choo, B. G. H., Kondrichin, I., Parinov, S., Emelyanov, A., Go, W., chang Toh, W., & Korzh, V. (2006). Zebrafish transgenic enhancer trap line database (zetrap). BMC Dev Biol *6*, 5.

[Chorley *et al.*, 2008] Chorley, B. N., Wang, X., Campbell, M. R., Pittman, G. S., Noureddine, M. A., & Bell, D. A. (2008). Discovery and verification

of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res *659*(1-2), 147–57.

[Coghlan *et al.*, 2005] Coghlan, A., Eichler, E. E., Oliver, S. G., Paterson, A. H., & Stein, L. (2005). Chromosome evolution in eukaryotes: a multi-kingdom perspective. Trends Genet. *21*(12), 673–82.

[Colgan & Manley, 1997] Colgan, D. F. & Manley, J. L. (1997). Mechanism and regulation of mrna polyadenylation. Genes Dev *11*(21), 2755–66.

[Coller *et al.*, 1998] Coller, J. M., Gray, N. K., & Wickens, M. P. (1998). mrna stabilization by poly(a) binding protein is independent of poly(a) and requires translation. Genes Dev *12*(20), 3226–35.

[Cooper *et al.*, 2006] Cooper, S., Trinklein, N., & Anton, E. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1 Genome Res.

[Cosma, 2002] Cosma, M. P. (2002). Ordered recruitment: gene-specific mechanism of transcription activation. Mol Cell *10*(2), 227–36.

[Crawford *et al.*, 2006] Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G., & Collins, F. S. (2006). Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). Genome Res *16*(1), 123–31.

[Crisponi *et al.*, 2004] Crisponi, L., Uda, M., Deiana, M., Loi, A., Nagaraja, R., Chiappe, F., Schlessinger, D., Cao, A., & Pilia, G. (2004). Foxl2 inactivation by a translocation 171 kb away: analysis of 500 kb of chromosome 3 for candidate long-range regulatory sequences. Genomics *83*(5), 757–64.

[Curwen *et al.*, 2004] Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J., & Clamp, M. (2004). The ensembl automatic gene annotation system.. Genome Res *14*(5), 942–50.

[Darai-Ramqvist *et al.*, 2008] Darai-Ramqvist, E., Sandlund, A., Müller, S., Klein, G., Imreh, S., & Kost-Alimova, M. (2008). Segmental duplications and evolutionary plasticity at tumor chromosome break-prone regions. Genome Res *18*(3), 370–9.

[de Kok *et al.*, 1996] de Kok, Y. J., Vossenaar, E. R., Cremers, C. W., Dahl, N., Laporte, J., Hu, L. J., Lacombe, D., Fischel-Ghodsian, N., Friedman, R. A.,

Parnes, L. S., Thorpe, P., Bitner-Glindzicz, M., Pander, H. J., Heilbronner, H., Graveline, J., den Dunnen, J. T., Brunner, H. G., Ropers, H. H., & Cremers, F. P. (1996). Identification of a hot spot for microdeletions in patients with x-linked deafness type 3 (dfn3) 900 kb proximal to the dfn3 gene pou3f4. Hum Mol Genet *5*(9), 1229–35.

[de la Calle-Mustienes *et al.*, 2005] de la Calle-Mustienes, E., Feijóo, C. G., Manzanares, M., Tena, J. J., Rodríguez-Seguel, E., Letizia, A., Allende, M. L., & Gómez-Skarmeta, J. L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate iroquois cluster gene deserts. Genome Res *15*(8), 1061–72.

[de Sá, 2007] de Sá, M. (2007). Rearrangements, http://iweb.langara.bc.ca/biology/mario/biol2330notes/biol2330chap8

[Dekker, 2003] Dekker, J. (2003). A closer look at long-range chromosomal interactions. Trends Biochem Sci *28*(6), 277–80.

[Dekker, 2008] Dekker, J. (2008). Gene regulation in the third dimension. Science *319*(5871), 1793–4.

[Dekker *et al.*, 2002] Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. Science *295*(5558), 1306–11.

[Dermitzakis *et al.*, 2002] Dermitzakis, E., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B., Flegel, V., Bucher, P., Jongeneel, C., & Antonarakis, S. (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature *420*(6915), 578–82. 0028-0836 (Print) Journal Article.

[Dillon *et al.*, 1997] Dillon, N., Trimborn, T., Strouboulis, J., Fraser, P., & Grosveld, F. (1997). The effect of distance on long-range chromatin interactions. Mol Cell *1*(1), 131–9. Example of the beta globin locus.

[Dostie *et al.*, 2006] Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., & Dekker, J. (2006). Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. Genome Res *16*(10), 1299–309.

[Duboule & Wilkins, 1998] Duboule, D. & Wilkins, A. S. (1998). The evolution of 'bricolage'. Trends Genet. *14*(2), 54–9.

[Durkin & Glover, 2007] Durkin, S. G. & Glover, T. W. (2007). Chromosome fragile sites. Annu Rev Genet *41*, 169–92.

[Eckhardt *et al.*, 2006] Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., & Beck, S. (2006). Dna methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet. *38*(12), 1378–85.

[Eckner *et al.*, 1994] Eckner, R., Ewen, M. E., Newsome, D., Gerdes, M., DeCaprio, J. A., Lawrence, J. B., & Livingston, D. M. (1994). Molecular cloning and functional analysis of the adenovirus e1a-associated 300-kd protein (p300) reveals a protein with properties of a transcriptional adaptor. Genes Dev *8*(8), 869–84.

[Ellingsen *et al.*, 2005] Ellingsen, S., Laplante, M. A., König, M., Kikuta, H., Furmanek, T., Hoivik, E. A., & Becker, T. S. (2005). Large-scale enhancer detection in the zebrafish genome. Development *132*(17), 3799–811.

[Ellington & Szostak, 1990] Ellington, A. D. & Szostak, J. W. (1990). In vitro selection of rna molecules that bind specific ligands. Nature *346*(6287), 818–22.

[Elnitski *et al.*, 2003] Elnitski, L., Hardison, R., Li, J., Yang, S., Kolbe, D., Eswara, P., O'Connor, M., Schwartz, S., Miller, W., & Chiaromonte, F. (2003). Distinguishing regulatory dna from neutral sites. Genome Res *13*(1), 64–72. 1088-9051 (Print) Journal Article.

[Engström *et al.*, 2007] Engström, P. G., Sui, S. J. H., Drivenes, O., Becker, T. S., & Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res *17*(12), 1898–908.

[Ermolaeva *et al.*, 2001] Ermolaeva, M. D., White, O., & Salzberg, S. L. (2001). Prediction of operons in microbial genomes. Nucleic Acids Res *29*(5), 1216–21.

[Ettwiller *et al.*, 2007] Ettwiller, L., Paten, B., Ramialison, M., Birney, E., & Wittbrodt, J. (2007). Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. Nat. Methods *4*(7), 563–5.

[Euskirchen *et al.*, 2004] Euskirchen, G., Royce, T. E., Bertone, P., Martone, R., Rinn, J. L., Nelson, F. K., Sayward, F., Luscombe, N. M., Miller, P., Gerstein, M., Weissman, S., & Snyder, M. (2004). Creb binds to multiple loci on human chromosome 22. Mol Cell Biol *24*(9), 3804–14.

[Farré *et al.*, 2007] Farré, D., Bellora, N., Mularoni, L., Messeguer, X., & Albà, M. (2007). Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol *8*(7), R140.

[Farrell *et al.*, 2002] Farrell, C. M., West, A. G., & Felsenfeld, G. (2002). Conserved ctcf insulator elements flank the mouse and human beta-globin loci. Mol Cell Biol *22*(11), 3820–31.

[Ferretti *et al.*, 2007] Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F., & Blanchette, M. (2007). Premod: a database of genome-wide mammalian cis-regulatory module predictions. Nucleic Acids Res *35*(Database issue), D122–6.

[Fitch & Margoliash, 1967] Fitch, W. M. & Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*(760), 279–84.

[Flicek *et al.*, 2008] Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., & Searle, S. (2008). Ensembl 2008. Nucleic Acids Res *36*(Database issue), D707–14.

[Flint *et al.*, 2001] Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R. J., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K. C., McMorrow, T., Frampton, J., Alter, B. P., Frischauf, A. M., & Higgs, D. R. (2001). Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. Hum Mol Genet *10*(4), 371–82.

[Fondell *et al.*, 1996] Fondell, J. D., Ge, H., & Roeder, R. G. (1996). Ligand induction of a transcriptionally active thyroid hormone receptor coactivator complex. Proc Natl Acad Sci USA *93*(16), 8329–33.

[Fraser *et al.*, 2009] Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M. A., Hayashizaki, Y., Blanchette, M., & Dostie, J. (2009). Chromatin conformation signatures of cellular differentiation. Genome Biol. *10*(4), R37.

[Fried & Crothers, 1981] Fried, M. & Crothers, D. M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Res *9*(23), 6505–25.

[Frith *et al.*, 2001] Frith, M. C., Hansen, U., & Weng, Z. (2001). Detection of cis-element clusters in higher eukaryotic dna. Bioinformatics *17*(10), 878–89.

[Frith *et al.*, 2003] Frith, M. C., Li, M. C., & Weng, Z. (2003). Cluster-buster: Finding dense clusters of motifs in dna sequences.. Nucleic Acids Res *31*(13), 3666–8.

[Frommer *et al.*, 1992] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., & Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. Proc Natl Acad Sci USA *89*(5), 1827–31.

[Fullwood *et al.*, 2009] Fullwood, M. J., Wei, C.-L., Liu, E. T., & Ruan, Y. (2009). Next-generation dna sequencing of paired-end tags (pet) for transcriptome and genome analyses. Genome Res *19*(4), 521–32.

[Gardiner-Garden & Frommer, 1987] Gardiner-Garden, M. & Frommer, M. (1987). Cpg islands in vertebrate genomes. J Mol Biol *196*(2), 261–82.

[Ge *et al.*, 2009] Ge, B., Pokholok, D. K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D. J., Le, J., Koka, V., Lam, K. C. L., Gagné, V., Dias, J., Hoberman, R., Montpetit, A., Joly, M.-M., Harvey, E. J., Sinnett, D., Beaulieu, P., Hamon, R., Graziani, A., Dewar, K., Harmsen, E., Majewski, J., Göring, H. H. H., Naumova, A. K., Blanchette, M., Gunderson, K. L., & Pastinen, T. (2009). Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat. Genet. *41*(11), 1216–22.

[Geyer & Corces, 1992] Geyer, P. K. & Corces, V. G. (1992). Dna position-specific repression of transcription by a drosophila zinc finger protein. Genes Dev *6*(10), 1865–73.

[Gibbs *et al.*, 2004] Gibbs, R., Weinstock, G., Metzker, M., Muzny, D., Sodergren, E., Scherer, S., Scott, G., Steffen, D., Worley, K., Burch, P., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R., Adams, M., Amanatides, P., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C., Nguyen, T., Pfannkoch, C., Sitter, C., Sutton, G., Venter, J., Woodage, T., Smith, D., Lee, H., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R., Dunn, D., Green, E., Blakesley, R., Bouffard, G., Jong, P. D., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W., Havlak, P., Chen, R., Durbin, K., Egan, A., Ren, Y., Song, X., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K., Cooney, A., D'Souza, L., Martin, K., Wu, J., Gonzalez-Garay, M., Jackson, A., Kalafus, K., McLeod, M., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D., Zhang, Z., Bailey, J., Eichler, E., Tuzun, E., Birney, E., Mongin, E., Ureta-Vidal, A., Woodwark, C., Zdobnov, E., Bork, P., Suyama, M., Torrents, D., Alexandersson, M., Trask, B., Young, J., Huang, H., Wang, H., Xing, H., Daniels, S., Gietzen, D., Schmidt, J., Stevens, K., Vitt, U., Wingrove, J., Camara, F., Alba, M. M., Abril, J., Guigo, R., Smit, A., Dubchak, I., Rubin, E., Couronne, O., Poliakov, A., Hubner, N., Ganten, D., Goesele, C., Hummel, O., Kreitler, T., Lee, Y., Monti, J., Schulz, H., Zimdahl, H., Himmelbauer, H., Lehrach, H., Jacob, H., Bromberg, S., Gullings-Handley, J., Jensen-Seaman, M., Kwitek, A., Lazar, J., Pasko, D., Tonellato, P., Twigger, S., Ponting, C., Duarte, J., Rice, S., Goodstadt, L., Beatson, S., Emes, R., Winter, E., Webber, C., Brandt, P., Nyakatura, G., Adetobi, M., Chiaromonte, F., Elnitski, L., Eswara, P., Hardison, R., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutenko, A., Riemer, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpaintner, K., Andrews, T., Caccamo, M., Clamp, M., Clarke, L., Curwen, V., Durbin, R., Eyras, E., Searle, S., Cooper, G., Batzoglou, S., Brudno, M., Sidow, A., Stone, E., Venter, J., Payseur, B., Bourque, G., Lopez-Otin, C., Puente, X., Chakrabarti, K., Chatterji, S., Dewey, C., Pachter, L., Bray, N., Yap, V., Caspi, A., Tesler, G., Pevzner, P., Haussler, D., Roskin, K., Baertsch, R., Clawson, H., Furey, T., Hinrichs, A., Karolchik, D., Kent, W., Rosenbloom,

K., Trumbower, H., Weirauch, M., Cooper, D., Stenson, P., Ma, B., Brent, M., Arumugam, M., Shteynberg, D., Copley, R., Taylor, M., Riethman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S., & Collins, F. (2004). Genome sequence of the brown norway rat yields insights into mammalian evolution. Nature *428*(6982), 493–521. 1476-4687 Journal Article.

[Giraldo & Montoliu, 2001] Giraldo, P. & Montoliu, L. (2001). Size matters: use of yacs, bacs and pacs in transgenic animals. Transgenic Res *10*(2), 83–103.

[Gómez-Skarmeta *et al.*, 2006] Gómez-Skarmeta, J. L., Lenhard, B., & Becker, T. S. (2006). New technologies, new findings, and new concepts in the study of vertebrate cis-regulatory sequences. Dev Dyn *235*(4), 870–85.

[Gonzalez *et al.*, 2007] Gonzalez, F., Duboule, D., & Spitz, F. (2007). Transgenic analysis of hoxd gene regulation during digit development. Dev. Biol.

[Goode *et al.*, 2005] Goode, D. K., Snell, P., Smith, S. F., Cooke, J. E., & Elgar, G. (2005). Highly conserved regulatory elements around the shh gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. Genomics *86*(2), 172–81.

[Grabher & Wittbrodt, 2007] Grabher, C. & Wittbrodt, J. (2007). Meganuclease and transposon mediated transgenesis in medaka. Genome Biol. *8 Suppl 1*, S10.

[Grad *et al.*, 2004] Grad, Y. H., Roth, F. P., Halfon, M. S., & Church, G. M. (2004). Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in drosophila melanogaster and d.pseudoobscura. Bioinformatics *20*(16), 2738–50.

[Gross & Garrard, 1988] Gross, D. S. & Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. Annu Rev Biochem *57*, 159–97.

[Gu *et al.*, 2008] Gu, W., Zhang, F., & Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. PathoGenetics *1*(1), 4.

[Gumucio *et al.*, 1992] Gumucio, D. L., Heilstedt-Williamson, H., Gray, T. A., Tarlé, S. A., Shelton, D. A., Tagle, D. A., Slightom, J. L., Goodman, M., & Collins, F. S. (1992). Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. Mol Cell Biol *12*(11), 4919–29.

[Hampsey & Reinberg, 1999] Hampsey, M. & Reinberg, D. (1999). Rna polymerase ii as a control panel for multiple coactivator complexes. Curr Opin Genet Dev *9*(2), 132–9.

[Hannenhalli & Levy, 2002] Hannenhalli, S. & Levy, S. (2002). Predicting transcription factor synergism. Nucleic Acids Res *30*(19), 4278–84.

[Harbison *et al.*, 2004] Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., & Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome.. Nature *431*(7004), 99–104.

[Haudry *et al.*, 2008] Haudry, Y., Berube, H., Letunic, I., Weeber, P.-D., Gagneur, J., Girardot, C., Kapushesky, M., Arendt, D., Bork, P., Brazma, A., Furlong, E. E. M., Wittbrodt, J., & Henrich, T. (2008). 4dxpress: a database for cross-species expression pattern comparisons. Nucleic Acids Res *36*(Database issue), D847–53.

[Havlak *et al.*, 2004] Havlak, P., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X.-Z., Weinstock, G. M., & Gibbs, R. A. (2004). The atlas genome assembly system. Genome Res *14*(4), 721–32.

[Heintz, 2001] Heintz, N. (2001). Bac to the future: the use of bac transgenic mice for neuroscience research. Nat Rev Neurosci *2*(12), 861–70.

[Hillier *et al.*, 2004] Hillier, L., Miller, W., Birney, E., Warren, W., Hardison, R., Ponting, C., Bork, P., Burt, D., Groenen, M., Delany, M., Dodgson, J., Chinwalla, A., Cliften, P., Clifton, S., Delehaunty, K., Fronick, C., Fulton, R., Graves, T., Kremitzki, C., Layman, D., Magrini, V., McPherson, J., Miner, T., Minx, P., Nash, W., Nhan, M., Nelson, J., Oddy, L., Pohl, C., Randall-Maher, J., Smith, S., Wallis, J., Yang, S., Romanov, M., Rondelli, C., Paton, B., Smith, J., Morrice, D., Daniels, L., Tempest, H., Robertson, L., Masabanda, J., Griffin, D., Vignal, A., Fillon, V., Jacobbson, L., Kerje, S., Andersson, L., Crooijmans, R., Aerts, J., van der Poel, J., Ellegren, H., Caldwell, R., Hubbard, S., Grafham, D., Kierzek, A., McLaren, S., Overton, I., Arakawa, H., Beattie, K., Bezzubov, Y., Boardman, P., Bonfield, J., Croning, M., Davies, R., Francis, M., Humphray, S., Scott, C., Taylor, R., Tickle, C., Brown, W., Rogers, J., Buerstedde, J., Wilson, S., Stubbs, L., Ovcharenko, I., Gordon, L., Lucas, S., Miller, M., Inoko,

H., Shiina, T., Kaufman, J., Salomonsen, J., Skjoedt, K., Wong, G., Wang, J., Liu, B., Wang, J., Yu, J., Yang, H., Nefedov, M., Koriabine, M., Dejong, P., Goodstadt, L., Webber, C., Dickens, N., Letunic, I., Suyama, M., Torrents, D., von Mering, C., Zdobnov, E., Makova, K., Nekrutenko, A., Elnitski, L., Eswara, P., King, D., Yang, S., Tyekucheva, S., Radakrishnan, A., Harris, R., Chiaromonte, F., Taylor, J., He, J., Rijnkels, M., Griffiths-Jones, S., Ureta-Vidal, A., Hoffman, M., Severin, J., Searle, S., Law, A., Speed, D., Waddington, D., Cheng, Z., Tuzun, E., Eichler, E., Bao, Z., Flicek, P., Shteynberg, D., Brent, M., Bye, J., Huckle, E., Chatterji, S., Dewey, C., Pachter, L., Kouranov, A., Mourelatos, Z., Hatzigeorgiou, A., Paterson, A., Ivarie, R., Brandstrom, M., Axelsson, E., Backstrom, N., Berlin, S., Webster, M., Pourquie, O., Reymond, A., Ucla, C., Antonarakis, S., Long, M., Emerson, J., Betran, E., Dupanloup, I., Kaessmann, H., Hinrichs, A., Bejerano, G., Furey, T., Harte, R., Raney, B., Siepel, A., Kent, W., Haussler, D., Eyras, E., Castelo, R., Abril, J., Castellano, S., Camara, F., Parra, G., Guigo, R., Bourque, G., Tesler, G., Pevzner, P., Smit, A., Fulton, L., Mardis, E., & Wilson, R. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*(7018), 695–716. 1476-4687 Journal Article.

[Howard & Davidson, 2004] Howard, M. L. & Davidson, E. H. (2004). cis-regulatory control circuits in development.. Dev. Biol. *271*(1), 109–18.

[Hubbard *et al.*, 2009] Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., & Flicek, P. (2009). Ensembl 2009. Nucleic Acids Res *37*(Database issue), D690–7.

[Hubbard *et al.*, 2007] Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp,

C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., & Birney, E. (2007). Ensembl 2007. Nucleic Acids Res *35*(Database issue), D610–7.

[Huynen *et al.*, 2000] Huynen, M., Snel, B., Lathe, W., & Bork, P. (2000). Exploitation of gene context. Current opinion in structural biology *10*(3), 366–70.

[Iafrate *et al.*, 2004] Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., & Lee, C. (2004). Detection of large-scale variation in the human genome. Nat. Genet. *36*(9), 949–51.

[ichi Higashijima *et al.*, 2003] ichi Higashijima, S., Masino, M. A., Mandel, G., & Fetcho, J. R. (2003). Imaging neuronal activity during zebrafish behavior with a genetically encoded calcium indicator. J Neurophysiol *90*(6), 3986–97.

[Ioshikhes & Zhang, 2000] Ioshikhes, I. P. & Zhang, M. Q. (2000). Large-scale human promoter mapping using cpg islands. Nat. Genet. *26*(1), 61–3.

[Ishikawa, 2000] Ishikawa, Y. (2000). Medakafish as a model system for vertebrate developmental genetics. Bioessays *22*(5), 487–95.

[Iwamatsu, 2004] Iwamatsu, T. (2004). Stages of normal development in the medaka oryzias latipes. Mech Dev *121*(7-8), 605–18.

[Iyer *et al.*, 2001] Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., & Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. Nature *409*(6819), 533–8.

[Jaenisch & Bird, 2003] Jaenisch, R. & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat. Genet. *33 Suppl*, 245–54.

[Jaillon *et al.*, 2004] Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., Jaffe, D., Fisher, S., Lutfalla, G., Dossat, C., Segurens, B., Dasilva, C., Salanoubat, M., Levy, M., Boudet, N., Castellano, S., Anthouard, V., Jubin, C., Castelli, V., Katinka, M., Vacherie, B., Biémont, C., Skalli, Z.,

Cattolico, L., Poulain, J., Berardinis, V. D., Cruaud, C., Duprat, S., Brottier, P., Coutanceau, J.-P., Gouzy, J., Parra, G., Lardier, G., Chapple, C., McKernan, K. J., McEwan, P., Bosak, S., Kellis, M., Volff, J.-N., Guigó, R., Zody, M. C., Mesirov, J., Lindblad-Toh, K., Birren, B., Nusbaum, C., Kahn, D., Robinson-Rechavi, M., Laudet, V., Schachter, V., Quétier, F., Saurin, W., Scarpelli, C., Wincker, P., Lander, E. S., Weissenbach, J., & Crollius, H. R. (2004). Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype.. Nature *431*(7011), 946–57.

[Johansson *et al.*, 2003] Johansson, O., Alkema, W., Wasserman, W. W., & Lagergren, J. (2003). Identification of functional clusters of transcription factor binding motifs in genome sequences: the mscan algorithm. Bioinformatics *19 Suppl 1*, i169–76.

[Johnson *et al.*, 2003] Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., & Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. Science *302*(5653), 2141–4.

[Kadonaga, 2002] Kadonaga, J. T. (2002). The dpe, a core promoter element for transcription by rna polymerase ii. Exp Mol Med *34*(4), 259–64.

[Kasahara *et al.*, 2007] Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S.-I., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S., & Kohara, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. Nature *447*(7145), 714–9.

[Kasprzyk *et al.*, 2004] Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., & Birney, E. (2004). Ensmart: a generic system for fast and flexible access to biological data. Genome Res *14*(1), 160–9.

[Kehrer-Sawatzki & Cooper, 2008] Kehrer-Sawatzki, H. & Cooper, D. N. (2008). Molecular mechanisms of chromosomal rearrangement during primate evolution. Chromosome Res *16*(1), 41–56.

[Kel *et al.*, 2003] Kel, A. E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., & Wingender, E. (2003). Match: A tool for searching transcription factor binding sites in dna sequences. Nucleic Acids Res *31*(13), 3576–9.

[Kellum & Schedl, 1991] Kellum, R. & Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. Cell *64*(5), 941–50.

[Kent *et al.*, 2002] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at ucsc. Genome Res *12*(6), 996–1006.

[Khokha & Loots, 2005] Khokha, M. K. & Loots, G. G. (2005). Strategies for characterising cis-regulatory elements in xenopus. Briefings in functional genomics & proteomics *4*(1), 58–68.

[Kidd *et al.*, 2008] Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R., & Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*(7191), 56–64.

[Kikuta *et al.*, 2007] Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., Ghislain, J., Pezeron, G., Mourrain, P., Ellingsen, S., Oates, A. C., Thisse, C., Thisse, B., Foucher, I., Adolf, B., Geling, A., Lenhard, B., & Becker, T. S. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res *17*(5), 545–55.

[Kirkpatrick & Barton, 2006] Kirkpatrick, M. & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. Genetics *173*(1), 419–34.

[Kleinjan *et al.*, 2001] Kleinjan, D. A., Seawright, A., Schedl, A., Quinlan, R. A., Danes, S., & van Heyningen, V. (2001). Aniridia-associated translocations,

dnase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of pax6. Hum Mol Genet *10*(19), 2049–59.

[Kleinjan & van Heyningen, 2005] Kleinjan, D. A. & van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet *76*(1), 8–32.

[Kloosterman & Plasterk, 2006] Kloosterman, W. P. & Plasterk, R. H. A. (2006). The diverse functions of micrornas in animal development and disease. Dev Cell *11*(4), 441–50. [TC SYNTENY].

[Klug & Famulok, 1994] Klug, S. & Famulok, M. (1994). All you wanted to know about selex. Molecular biology reports.

[Korzh, 2007] Korzh, V. (2007). Transposons as tools for enhancer trap screens in vertebrates. Genome Biol. *8 Suppl 1*, S8.

[Kost-Alimova *et al.*, 2003] Kost-Alimova, M., Kiss, H., Fedorova, L., Yang, Y., Dumanski, J. P., Klein, G., & Imreh, S. (2003). Coincidence of synteny breakpoints with malignancy-related deletions on human chromosome 3. Proc Natl Acad Sci USA *100*(11), 6622–7.

[Kouzarides, 2007] Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*(4), 693–705.

[Lander *et al.*, 2001] Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R., Wilson, R., Hillier, L., McPherson, J., Marra, M., Mardis, E., Fulton, L., Chinwalla, A., Pepin, K., Gish, W., Chissoe, S., Wendl, M., Delehaunty, K., Miner, T., Delehaunty, A., Kramer, J., Cook, L., Fulton, R., Johnson, D., Minx, P., Clifton, S., Hawkins, T., Branscomb, E., Predki, P.,

Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R., Muzny, D., Scherer, S., Bouck, J., Sodergren, E., Worley, K., Rives, C., Gorrell, J., Metzker, M., Naylor, S., Kucherlapati, R., Nelson, D., Weinstock, G., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R., Federspiel, N., Abola, A., Proctor, M., Myers, R., Schmutz, J., Dickson, M., Grimwood, J., Cox, D., Olson, M., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G., Athanasiou, M., Schultz, R., Roe, B., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D., Burge, C., Cerutti, L., Chen, H., Church, D., Clamp, M., Copley, R., Doerks, T., Eddy, S., Eichler, E., Furey, T., Galagan, J., Gilbert, J., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L., Jones, T., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W., Kitts, P., Koonin, E., Korf, I., Kulp, D., Lancet, D., Lowe, T., McLysaght, A., Mikkelsen, T., Moran, J., Mulder, N., Pollara, V., Ponting, C., Schuler, G., Schultz, J., Slater, G., Smit, A., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y., Wolfe, K., Yang, S., Yeh, R., Collins, F., Guyer, M., Peterson, J., Felsenfeld, A., Wetterstrand, K., Patrinos, A., Morgan, M., de Jong, P., Catanese, J., Osoegawa, K., Shizuya, H., Choi, S., & Chen, Y. (2001). Initial sequencing and analysis of the human genome. Nature *409*(6822), 860–921. 0028-0836 (Print) Journal Article.

[Larkin *et al.*, 2009] Larkin, D. M., Pape, G., Donthu, R., Auvil, L., Welge, M., & Lewin, H. A. (2009). Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Res *19*(5), 770–7.

[Larkin *et al.*, 2003] Larkin, D. M., van der Wind, A. E., Rebeiz, M., Schweitzer, P. A., Bachman, S., Green, C., Wright, C. L., Campos, E. J., Benson, L. D., Edwards, J., Liu, L., Osoegawa, K., Womack, J. E., de Jong, P. J., & Lewin, H. A. (2003). A cattle-human comparative map built with cattle bac-ends and human genome sequence. Genome Res *13*(8), 1966–72.

[Lee *et al.*, 2006] Lee, A. P., Koh, E. G. L., Tay, A., Brenner, S., & Venkatesh, B. (2006). Highly conserved syntenic blocks at the vertebrate hox loci and conserved regulatory elements within and outside hox gene clusters. Proc Natl Acad Sci USA *103*(18), 6994–9.

[Lee *et al.*, 2008] Lee, J., Han, K., Meyer, T. J., Kim, H.-S., & Batzer, M. A. (2008). Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. PLoS ONE *3*(12), e4047.

[Lee *et al.*, 1993] Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. Cell *75*(5), 843–54.

[Lee *et al.*, 2002] Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., & Young, R. A. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. Science *298*(5594), 799–804.

[Leipoldt *et al.*, 2007] Leipoldt, M., Erdel, M., Bien-Willner, G. A., Smyk, M., Theurl, M., Yatsenko, S. A., Lupski, J. R., Lane, A. H., Shanske, A. L., Stankiewicz, P., & Scherer, G. (2007). Two novel translocation breakpoints upstream of sox9 define borders of the proximal and distal breakpoint cluster region in campomelic dysplasia. Clin Genet *71*(1), 67–75.

[Lenhard *et al.*, 2003] Lenhard, B., Sandelin, A., Mendoza, L., Engström, P., Jareborg, N., & Wasserman, W. W. (2003). Identification of conserved regulatory elements by comparative genome analysis. J Biol *2*(2), 13.

[Lettice *et al.*, 2003] Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., & de Graaff, E. (2003). A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet *12*(14), 1725–35.

[Lewis *et al.*, 2005] Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. Cell *120*(1), 15–20.

[Lieb *et al.*, 2001] Lieb, J., Liu, X., Botstein, D., & Brown, P. (2001). ... -specific binding of rap1 revealed by genome-wide maps of protein–dna association. Nat. Genet.

[Lifanov *et al.*, 2003] Lifanov, A. P., Makeev, V. J., Nazina, A. G., & Papatsenko, D. A. (2003). Homotypic regulatory clusters in drosophila. Genome Res *13*(4), 579–88.

[Lindblad-Toh *et al.*, 2005] Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., DeJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A. C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J.-P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S.,

Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A., & Lander, E. S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog.. Nature *438*(7069), 803–19.

[Liu *et al.*, 2001] Liu, Y., Bondarenko, V., Ninfa, A., & Studitsky, V. M. (2001). Dna supercoiling allows enhancer action over a large distance. Proc Natl Acad Sci USA *98*(26), 14883–8.

[Lodish *et al.*, 1995] Lodish, H., Berk, A., Baltimore, D., Zipursky, S. L., Matsudaira, P., & Darnell, J. (1995). Molecular cell biology.

[Loosli *et al.*, 2001] Loosli, F., Winkler, S., Burgtorf, C., Wurmbach, E., Ansorge, W., Henrich, T., Grabher, C., Arendt, D., Carl, M., Krone, A., Grzebisz, E., & Wittbrodt, J. (2001). Medaka eyeless is the key factor linking retinal determination and eye growth. Development *128*(20), 4035–44.

[Loots *et al.*, 2000] Loots, G., Locksley, R., Blankespoor, C., Wang, Z., Miller, W., Rubin, E., & Frazer, K. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science *288*(5463), 136–40. 0036-8075 (Print) Journal Article.

[Loots *et al.*, 2002] Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I., & Rubin, E. M. (2002). rvista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res *12*(5), 832–9.

[Lupski & Stankiewicz, 2005] Lupski, J. R. & Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet *1*(6), e49.

[Mackenzie *et al.*, 2004] Mackenzie, A., Miller, K. A., & Collinson, J. M. (2004). Is there a functional link between gene interdigitation and multi-species conservation of synteny blocks?. Bioessays *26*(11), 1217–24.

[Maizels, 2005] Maizels, N. (2005). Immunoglobulin gene diversification. Annu Rev Genet *39*, 23–46.

[Martone *et al.*, 2003] Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T. E., Luscombe, N. M., Rinn, J. L., Nelson, F. K., Miller, P., Gerstein, M., Weissman, S., & Snyder, M. (2003). Distribution of nf-kappab-binding sites across human chromosome 22. Proc Natl Acad Sci USA *100*(21), 12247–52.

[Maston *et al.*, 2006] Maston, G., Evans, S., & Green, M. (2006). Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet *7*, 29–59. 1527-8204 (Print) Journal article.

[Mattick & Makunin, 2006] Mattick, J. S. & Makunin, I. V. (2006). Non-coding rna. Hum Mol Genet *15 Spec No 1*, R17–29.

[Matys *et al.*, 2006] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., & Wingender, E. (2006). Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*(Database issue), D108–10.

[McFadden *et al.*, 2000] McFadden, D. G., Charité, J., Richardson, J. A., Srivastava, D., Firulli, A. B., & Olson, E. N. (2000). A gata-dependent right ventricular enhancer controls dhand transcription in the developing heart. Development *127*(24), 5331–41.

[Megason & Fraser, 2003] Megason, S. G. & Fraser, S. E. (2003). Digitizing life at the level of the cell: high-performance laser-scanning microscopy and image analysis for in toto imaging of development. Mech Dev *120*(11), 1407–20.

[Merla *et al.*, 2006] Merla, G., Howald, C., Henrichsen, C. N., Lyle, R., Wyss, C., Zabot, M.-T., Antonarakis, S. E., & Reymond, A. (2006). Submicroscopic deletion in patients with williams-beuren syndrome influences expression levels of the nonhemizygous flanking genes. Am J Hum Genet *79*(2), 332–41.

[Mikkelsen *et al.*, 2007a] Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., & Bernstein, B. E. (2007a). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*(7153), 553–60.

[Mikkelsen *et al.*, 2007b] Mikkelsen, T. S., Wakefield, M. J., Aken, B., Amemiya, C. T., Chang, J. L., Duke, S., Garber, M., Gentles, A. J., Goodstadt, L., Heger, A., Jurka, J., Kamal, M., Mauceli, E., Searle, S. M. J., Sharpe, T., Baker, M. L., Batzer, M. A., Benos, P. V., Belov, K., Clamp, M., Cook, A., Cuff, J., Das, R., Davidow, L., Deakin, J. E., Fazzari, M. J., Glass, J. L., Grabherr, M., Greally, J. M., Gu, W., Hore, T. A., Huttley, G. A., Kleber, M., Jirtle, R. L., Koina, E., Lee, J. T., Mahony, S., Marra, M. A., Miller, R. D., Nicholls, R. D., Oda, M., Papenfuss, A. T., Parra, Z. E., Pollock, D. D., Ray, D. A., Schein, J. E., Speed, T. P., Thompson, K., VandeBerg, J. L., Wade, C. M., Walker, J. A., Waters, P. D., Webber, C., Weidman, J. R., Xie, X., Zody, M. C., Platform, B. I. G. S., Team, B. I. W. G. A., Graves, J. A. M., Ponting, C. P., Breen, M., Samollow, P. B., Lander, E. S., & Lindblad-Toh, K. (2007b). Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. Nature *447*(7141), 167–77.

[Miller *et al.*, 2007] Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., Pond, S. L. K., Nekrutenko, A., Giardine, B., Harris, R. S., Tyekucheva, S., Diekhans, M., Pringle, T. H., Murphy, W. J., Lesk, A., Weinstock, G. M., Lindblad-Toh, K., Gibbs, R. A., Lander, E. S., Siepel, A., Haussler, D., & Kent, W. J. (2007). 28-way vertebrate alignment and conservation track in the ucsc genome browser. Genome Res *17*(12), 1797–808.

[Mongin *et al.*, 2009] Mongin, E., Dewar, K., & Blanchette, M. (2009). Long-range regulation is a major driving force in maintaining genome integrity. BMC Evol. Biol. *9*(1), 203.

[Monteilhet *et al.*, 1990] Monteilhet, C., Perrin, A., Thierry, A., Colleaux, L., & Dujon, B. (1990). Purification and characterization of the in vitro activity of i-sce i, a novel and highly specific endonuclease encoded by a group i intron. Nucleic Acids Res *18*(6), 1407–13.

[Müller *et al.*, 2002] Müller, F., Blader, P., & Strähle, U. (2002). Search for enhancers: teleost models in comparative genomic and transgenic analysis of cis regulatory elements. Bioessays *24*(6), 564–72.

[Müller *et al.*, 1999] Müller, F., Chang, B., Albert, S., Fischer, N., Tora, L., & Strähle, U. (1999). Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. Development *126*(10), 2103–16.

[Murphy *et al.*, 2005] Murphy, W. J., Larkin, D. M., van der Wind, A. E., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A., Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M., Womack, J. E., O'brien, S. J., Pevzner, P. A., & Lewin, H. A. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science *309*(5734), 613–7.

[Nadeau & Taylor, 1984] Nadeau, J. H. & Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci USA *81*(3), 814–8.

[Nagai *et al.*, 2001] Nagai, T., Sawano, A., Park, E. S., & Miyawaki, A. (2001). Circularly permuted green fluorescent proteins engineered to sense ca2+. Proc Natl Acad Sci USA *98*(6), 3197–202.

[Nelson *et al.*, 2004] Nelson, C. E., Hersh, B. M., & Carroll, S. B. (2004). The regulatory content of intergenic dna shapes genome architecture. Genome Biol. *5*(4), R25.

[Nobile *et al.*, 2002] Nobile, C., Toffolatti, L., Rizzi, F., Simionati, B., Nigro, V., Cardazzo, B., Patarnello, T., Valle, G., & Danieli, G. A. (2002). Analysis of 22 deletion breakpoints in dystrophin intron 49. Hum Genet *110*(5), 418–21.

[Nobrega *et al.*, 2003] Nobrega, M. A., Ovcharenko, I., Afzal, V., & Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*(5644), 413. ts.

[Nóbrega *et al.*, 2004] Nóbrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V., & Rubin, E. M. (2004). Megabase deletions of gene deserts result in viable mice. Nature *431*(7011), 988–93.

[Odom *et al.*, 2007] Odom, D., Dowell, R., Jacobsen, E., Gordon, W., Danford, T., Macisaac, K., Rolfe, P., Conboy, C., Gifford, D., & Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat. Genet.

[Orlando, 2000] Orlando, V. (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem Sci *25*(3), 99–104.

[Ovcharenko *et al.*, 2005] Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W., & Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. Genome Res *15*(1), 137–45.

[Parinov *et al.*, 2004] Parinov, S., Kondrichin, I., Korzh, V., & Emelyanov, A. (2004). Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes in vivo. Dev Dyn *231*(2), 449–59.

[Paten *et al.*, 2009] Paten, B., Herrero, J., Beal, K., & Birney, E. (2009). Sequence progressive alignment, a framework for practical large-scale probabilistic consistency alignment. Bioinformatics *25*(3), 295–301.

[Paten *et al.*, 2008] Paten, B., Herrero, J., Beal, K., Fitzgerald, S., & Birney, E. (2008). Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res *18*(11), 1814–28.

[Peng *et al.*, 2006] Peng, Q., Pevzner, P. A., & Tesler, G. (2006). The fragile breakage versus random breakage models of chromosome evolution. PLoS Comput Biol *2*(2), e14.

[Pennacchio *et al.*, 2006] Pennacchio, L. A., Ahituv, N., Moses, A., Prabhakar, S., Nobrega, M., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K., Plajzer-Frick, I., Akiyama, J., Val, S. D., Afzal, V., Black, B., Couronne, O., Eisen, M., Visel, A., & Rubin, E. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*(7118), 499–502.

[Pertz *et al.*, 2006] Pertz, O., Hodgson, L., Klemke, R. L., & Hahn, K. M. (2006). Spatiotemporal dynamics of rhoa activity in migrating cells. Nature *440*(7087), 1069–72.

[Pevzner & Tesler, 2003] Pevzner, P. & Tesler, G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci USA *100*(13), 7672–7.

[Philippakis *et al.*, 2005] Philippakis, A. A., He, F. S., & Bulyk, M. L. (2005). Modulefinder: a tool for computational discovery of cis regulatory modules. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing , 519–30.

[Plessy *et al.*, 2005] Plessy, C., Dickmeis, T., Chalmel, F., & Strähle, U. (2005). Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. Trends Genet. *21*(4), 207–10.

[Pollock & Treisman, 1990] Pollock, R. & Treisman, R. (1990). A sensitive method for the determination of protein-dna binding specificities. Nucleic Acids Res *18*(21), 6197–204.

[Pop *et al.*, 2004] Pop, R., Conz, C., Lindenberg, K. S., Blesson, S., Schmalenberger, B., Briault, S., Pfeifer, D., & Scherer, G. (2004). Screening of the 1 mb sox9 5' control region by array cgh identifies a large deletion in a case of campomelic dysplasia with xy sex reversal. J Med Genet *41*(4), e47.

[Ptashne, 1988] Ptashne, M. (1988). How eukaryotic transcriptional activators work. Nature *335*(6192), 683–9.

[Ptashne & Gann, 1997] Ptashne, M. & Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*(6625), 569–77.

[Rajewsky *et al.*, 2002] Rajewsky, N., Vergassola, M., Gaul, U., & Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. BMC Bioinformatics *3*, 30.

[Rastegar, 2002] Rastegar, S. (2002). A floor plate enhancer of the zebrafish netrin1 gene requires cyclops (nodal) signalling and the winged helix transcription factor foxa2. Dev. Biol. *252*(1), 1–14.

[Rembold *et al.*, 2006] Rembold, M., Lahiri, K., Foulkes, N. S., & Wittbrodt, J. (2006). Transgenesis in fish: efficient selection of transgenic fish by co-injection with a fluorescent reporter construct. Nature protocols *1*(3), 1133–9.

[Ren *et al.*, 2000] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., & Young, R. A. (2000). Genome-wide location and function of dna binding proteins. Science *290*(5500), 2306–9.

[Roh *et al.*, 2006] Roh, T.-Y., Cuddapah, S., Cui, K., & Zhao, K. (2006). The genomic landscape of histone modifications in human t cells. Proc Natl Acad Sci USA *103*(43), 15782–7.

[Rollins *et al.*, 2006] Rollins, R. A., Haghighi, F., Edwards, J. R., Das, R., Zhang, M. Q., Ju, J., & Bestor, T. H. (2006). Large-scale structure of genomic methylation patterns. Genome Res *16*(2), 157–63.

[Roskin *et al.*, 2003] Roskin, K., Diekhans, M., & Haussler, D. (2003). Scoring two-species local alignments to try to statistically separate neutrally evolving from .... Proceedings of the seventh annual international ....

[Roth *et al.*, 1985] Roth, D. B., Porter, T. N., & Wilson, J. H. (1985). Mechanisms of nonhomologous recombination in mammalian cells. Mol Cell Biol *5*(10), 2599–607.

[Roth & Wilson, 1986] Roth, D. B. & Wilson, J. H. (1986). Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. Mol Cell Biol *6*(12), 4295–304.

[Ruiz-Herrera *et al.*, 2006] Ruiz-Herrera, A., Castresana, J., & Robinson, T. J. (2006). Is mammalian chromosomal evolution driven by regions of genome fragility?. Genome Biol. *7*(12), R115.

[Ruiz-Herrera & Robinson, 2008] Ruiz-Herrera, A. & Robinson, T. J. (2008). Evolutionary plasticity and cancer breakpoints in human chromosome 3. Bioessays *30*(11-12), 1126–37.

[Sagai *et al.*, 2005] Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., & Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific shh expression and truncation of the mouse limb. Development *132*(4), 797–803.

[Saltzman *et al.*, 2008] Saltzman, A. L., Kim, Y. K., Pan, Q., Fagnani, M. M., Maquat, L. E., & Blencowe, B. J. (2008). Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mrna decay. Mol Cell Biol *28*(13), 4320–30.

[Sandelin *et al.*, 2003] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., & Lenhard, B. (2003). Jaspar: an open-access database for eukaryotic transcription factor binding profiles.. Nucleic Acids Res *32*(Database issue), D91–4.

[Sandelin *et al.*, 2004] Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., & Lenhard, B. (2004). Arrays of

ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics *5*(1), 99.

[Sandelin *et al.*, 2007] Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., & Hume, D. A. (2007). Mammalian rna polymerase ii core promoters: insights from genome-wide studies. Nat Rev Genet *8*(6), 424–36.

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

[Sankoff & Nadeau, 2003] Sankoff, D. & Nadeau, J. H. (2003). Chromosome rearrangements in evolution: From gene order to genome sequence and back. Proc Natl Acad Sci USA *100*(20), 11188–9.

[Sankoff & Trinh, 2005] Sankoff, D. & Trinh, P. (2005). Chromosomal breakpoint reuse in genome sequence rearrangement. J. Comput. Biol. *12*(6), 812–21.

[Schones *et al.*, 2007] Schones, D. E., Smith, A. D., & Zhang, M. Q. (2007). Statistical significance of cis-regulatory modules. BMC Bioinformatics *8*, 19.

[Schroeder *et al.*, 2004] Schroeder, M. D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E. D., & Gaul, U. (2004). Transcriptional control in the segmentation gene network of drosophila. PLoS Biol *2*(9), E271.

[Schwartz *et al.*, 2006] Schwartz, M., Zlotorynski, E., & Kerem, B. (2006). The molecular basis of common and rare fragile sites. Cancer Lett *232*(1), 13–26.

[Schwartz *et al.*, 2003] Schwartz, S., Kent, W., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., & Miller, W. (2003). Human-mouse alignments with blastz. Genome Res *13*(1), 103–7. 1088-9051 Journal Article.

[Segal *et al.*, 2006] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z., & Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*(7104), 772–8.

[Sequencing *et al.*, 2009] Sequencing, B. G., Consortium, A., Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., Adelson, D. L., Eichler, E. E., Elnitski, L., Guigó, R., Hamernik, D. L., Kappes, S. M., Lewin, H. A., Lynn, D. J., Nicholas, F. W., Reymond, A., Rijnkels, M., Skow, L. C., Zdobnov, E. M., Schook, L., Womack, J., Alioto, T., Antonarakis, S. E., Astashyn, A., Chapple, C. E., Chen, H.-C., Chrast, J., Câmara, F., Ermolaeva,

O., Henrichsen, C. N., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Kokocinski, F., Landrum, M., Maglott, D., Pruitt, K., Sapojnikov, V., Searle, S. M., Solovyev, V., Souvorov, A., Ucla, C., Wyss, C., Anzola, J. M., Gerlach, D., Elhaik, E., Graur, D., Reese, J. T., Edgar, R. C., McEwan, J. C., Payne, G. M., Raison, J. M., Junier, T., Kriventseva, E. V., Eyras, E., Plass, M., Donthu, R., Larkin, D. M., Reecy, J., Yang, M. Q., Chen, L., Cheng, Z., Chitko-McKown, C. G., Liu, G. E., Matukumalli, L. K., Song, J., Zhu, B., Bradley, D. G., Brinkman, F. S. L., Lau, L. P. L., Whiteside, M. D., Walker, A., Wheeler, T. T., Casey, T., German, J. B., Lemay, D. G., Maqbool, N. J., Molenaar, A. J., Seo, S., Stothard, P., Baldwin, C. L., Baxter, R., Brinkmeyer-Langford, C. L., Brown, W. C., Childers, C. P., Connelley, T., Ellis, S. A., Fritz, K., Glass, E. J., Herzig, C. T. A., Iivanainen, A., Lahmers, K. K., Bennett, A. K., Dickens, C. M., Gilbert, J. G. R., Hagen, D. E., Salih, H., Aerts, J., Caetano, A. R., Dalrymple, B., Garcia, J. F., Gill, C. A., Hiendleder, S. G., Memili, E., Spurlock, D., Williams, J. L., Alexander, L., Brownstein, M. J., Guan, L., Holt, R. A., Jones, S. J. M., Marra, M. A., Moore, R., Moore, S. S., Roberts, A., Taniguchi, M., Waterman, R. C., Chacko, J., Chandrabose, M. M., Cree, A., Dao, M. D., Dinh, H. H., Gabisi, R. A., Hines, S., Hume, J., Jhangiani, S. N., Joshi, V., Kovar, C. L., Lewis, L. R., Liu, Y.-S., Lopez, J., Morgan, M. B., Nguyen, N. B., Okwuonu, G. O., Ruiz, S. J., Santibanez, J., Wright, R. A., Buhay, C., Ding, Y., Dugan-Rocha, S., Herdandez, J., Holder, M., Sabo, A., Egan, A., Goodell, J., Wilczek-Boney, K., Fowler, G. R., Hitchens, M. E., Lozado, R. J., Moen, C., Steffen, D., Warren, J. T., Zhang, J., Chiu, R., Schein, J. E., Durbin, K. J., Havlak, P., Jiang, H., Liu, Y., Qin, X., Ren, Y., Shen, Y., Song, H., Bell, S. N., Davis, C., Johnson, A. J., Lee, S., Nazareth, L. V., Patel, B. M., Pu, L.-L., Vattathil, S., Williams, R. L., Curry, S., Hamilton, C., Sodergren, E., Wheeler, D. A., Barris, W., Bennett, G. L., Eggen, A., Green, R. D., Harhay, G. P., Hobbs, M., Jann, O., Keele, J. W., Kent, M. P., Lien, S., McKay, S. D., McWilliam, S., Ratnakumar, A., Schnabel, R. D., Smith, T., Snelling, W. M., Sonstegard, T. S., Stone, R. T., Sugimoto, Y., Takasuga, A., Taylor, J. F., Tassell, C. P. V., Macneil, M. D., Abatepaulo, A. R. R., Abbey, C. A., Ahola, V., Almeida, I. G., Amadio, A. F., Anatriello, E., Bahadue, S. M., Biase, F. H., Boldt, C. R., Carroll, J. A., Carvalho, W. A., Cervelatti, E. P., Chacko, E., Chapin, J. E., Cheng, Y., Choi, J., Colley, A. J., de Campos, T. A., Donato, M. D., de Miranda Santos, I. K. F., de Oliveira, C. J. F., Deobald, H., Devinoy, E., Donohue, K. E., Dovc, P., Eberlein, A., Fitzsimmons, C. J., Franzin, A. M., Garcia, G. R., Genini, S., Gladney, C. J., Grant, J. R., Greaser, M. L., Green, J. A., Hadsell, D. L., Hakimov, H. A.,

Halgren, R., Harrow, J. L., Hart, E. A., Hastings, N., Hernandez, M., Hu, Z.-L., Ingham, A., Iso-Touru, T., Jamis, C., Jensen, K., Kapetis, D., Kerr, T., Khalil, S. S., Khatib, H., Kolbehdari, D., Kumar, C. G., Kumar, D., Leach, R., Lee, J. C.-M., Li, C., Logan, K. M., Malinverni, R., Marques, E., Martin, W. F., Martins, N. F., Maruyama, S. R., Mazza, R., McLean, K. L., Medrano, J. F., Moreno, B. T., Moré, D. D., Muntean, C. T., Nandakumar, H. P., Nogueira, M. F. G., Olsaker, I., Pant, S. D., Panzitta, F., Pastor, R. C. P., Poli, M. A., Poslusny, N., Rachagani, S., Ranganathan, S., Razpet, A., Riggs, P. K., Rincon, G., Rodriguez-Osorio, N., Rodriguez-Zas, S. L., Romero, N. E., Rosenwald, A., Sando, L., Schmutz, S. M., Shen, L., Sherman, L., Southey, B. R., Lutzow, Y. S., Sweedler, J. V., Tammen, I., Telugu, B. P. V. L., Urbanski, J. M., Utsunomiya, Y. T., Verschoor, C. P., Waardenberg, A. J., Wang, Z., Ward, R., Weikard, R., Welsh, T. H., White, S. N., Wilming, L. G., Wunderlich, K. R., Yang, J., & Zhao, F.-Q. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science *324*(5926), 522–8.

[Sharan *et al.*, 2003] Sharan, R., Ovcharenko, I., Ben-Hur, A., & Karp, R. M. (2003). Creme: a framework for identifying cis-regulatory modules in human-mouse conserved segments. Bioinformatics *19 Suppl 1*, i283–91.

[Shaw & Lupski, 2004] Shaw, C. J. & Lupski, J. R. (2004). Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet *13 Spec No 1*, R57–64.

[Shiraki *et al.*, 2003] Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., & Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA *100*(26), 15776–81.

[Siepel *et al.*, 2005] Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., Richards, S., Weinstock, G., Wilson, R., Gibbs, R., Kent, W., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*(8), 1034–50. 1088-9051 (Print) Journal Article.

[Simonis *et al.*, 2006] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., & de Laat, W. (2006). Nuclear organization of

active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). Nat. Genet. *38*(11), 1348–54.

[Sinha *et al.*, 2004] Sinha, S., Schroeder, M. D., Unnerstall, U., Gaul, U., & Siggia, E. D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in drosophila. BMC Bioinformatics *5*, 129.

[Smale & Kadonaga, 2003] Smale, S. T. & Kadonaga, J. T. (2003). The rna polymerase ii core promoter. Annu Rev Biochem *72*, 449–79.

[Smith *et al.*, 2007] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Consortium, O., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. *25*(11), 1251–5.

[Sonenberg & Hinnebusch, 2009] Sonenberg, N. & Hinnebusch, A. G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell *136*(4), 731–45.

[Spitz *et al.*, 2003] Spitz, F., Gonzalez, F., & Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the hoxd cluster. Cell *113*(3), 405–17.

[Spitz *et al.*, 2001] Spitz, F., Gonzalez, F., Peichel, C., Vogt, T. F., Duboule, D., & Zákány, J. (2001). Large scale transgenic and cluster deletion analysis of the hoxd complex separate an ancestral regulatory module from evolutionary innovations. Genes Dev *15*(17), 2209–14.

[Sprague *et al.*, 2006] Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D., Mani, P., Ramachandran, S., Schaper, K., Segerdell, E., Song, P., Sprunger, B., Taylor, S., Slyke, C. V., & Westerfield, M. (2006). The zebrafish information network: the zebrafish model organism database. Nucleic Acids Res *34*(Database issue), D581–5. 1362-4962 (Electronic) Journal Article.

[Srivastava *et al.*, 2007] Srivastava, J., Barber, D. L., & Jacobson, M. P. (2007). Intracellular ph sensors: design principles and functional significance. Physiology (Bethesda, Md) *22*, 30–9.

[Stankiewicz *et al.*, 2003] Stankiewicz, P., Shaw, C. J., Dapper, J. D., Wakui, K., Shaffer, L. G., Withers, M., Elizondo, L., Park, S.-S., & Lupski, J. R. (2003). Genome architecture catalyzes nonrecurrent chromosomal rearrangements. Am J Hum Genet *72*(5), 1101–16.

[Steidl *et al.*, 2007] Steidl, U., Steidl, C., Ebralidze, A., Chapuy, B., Han, H.-J., Will, B., Rosenbauer, F., Becker, A., Wagner, K., Koschmieder, S., Kobayashi, S., Costa, D. B., Schulz, T., O'Brien, K. B., Verhaak, R. G. W., Delwel, R., Haase, D., Trümper, L., Krauter, J., Kohwi-Shigematsu, T., Griesinger, F., & Tenen, D. G. (2007). A distal single nucleotide polymorphism alters long-range regulation of the pu.1 gene in acute myeloid leukemia. J Clin Invest *117*(9), 2611–20.

[Stormo, 2000] Stormo, G. D. (2000). Dna binding sites: representation and discovery. Bioinformatics *16*(1), 16–23.

[Su *et al.*, 2004] Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M., Walker, J., & Hogenesch, J. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*(16), 6062–7. 0027-8424 (Print) Journal Article.

[Sun & Elgin, 1999] Sun, F. L. & Elgin, S. C. (1999). Putting boundaries on silence. Cell *99*(5), 459–62.

[Sun *et al.*, 2008] Sun, H., Skogerbø, G., Wang, Z., Liu, W., & Li, Y. (2008). Structural relationships between highly conserved elements and genes in vertebrate genomes. PLoS ONE *3*(11), e3727.

[Tagle *et al.*, 1988] Tagle, D., Koop, B., Goodman, M., Slightom, J., Hess, D., & Jones, R. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol *203*(2), 439–55. 0022-2836 (Print) Journal Article.

[Tan *et al.*, 2007] Tan, S., Guo, J., Huang, Q., Chen, X., Li-Ling, J., Li, Q., & Ma, F. (2007). Retained introns increase putative microrna targets within 3' utrs of human mrna. FEBS Lett *581*(6), 1081–6.

[Tanaka *et al.*, 2001] Tanaka, M., Kinoshita, M., Kobayashi, D., & Nagahama, Y. (2001). Establishment of medaka (oryzias latipes) transgenic lines with the

expression of green fluorescent protein fluorescence exclusively in germ cells: a useful model to monitor germ cells in a live vertebrate. Proc Natl Acad Sci USA *98*(5), 2544–9.

[Thomas & Chiang, 2006] Thomas, M. C. & Chiang, C.-M. (2006). The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol *41*(3), 105–78.

[Toffolatti *et al.*, 2002] Toffolatti, L., Cardazzo, B., Nobile, C., Danieli, G. A., Gualandi, F., Muntoni, F., Abbs, S., Zanetti, P., Angelini, C., Ferlini, A., Fanin, M., & Patarnello, T. (2002). Investigating the mechanism of chromosomal deletion: characterization of 39 deletion breakpoints in introns 47 and 48 of the human dystrophin gene. Genomics *80*(5), 523–30.

[Trembath *et al.*, 2004] Trembath, D. G., Semina, E. V., Jones, D. H., Patil, S. R., Qian, Q., Amendt, B. A., Russo, A. F., & Murray, J. C. (2004). Analysis of two translocation breakpoints and identification of a negative regulatory element in patients with rieger's syndrome. Birth Defects Res Part A Clin Mol Teratol *70*(2), 82–91.

[Trinklein *et al.*, 2004] Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., & Myers, R. M. (2004). An abundance of bidirectional promoters in the human genome. Genome Res *14*(1), 62–6.

[Tsien, 1998] Tsien, R. Y. (1998). The green fluorescent protein. Annu Rev Biochem *67*, 509–44.

[Vangjeli *et al.*, 2010] Vangjeli, C., Clarke, N., Quinn, U., Dicker, P., Tighe, O., Ho, C., O'Brien, E., & Stanton, A. V. (2010). Confirmation that the renin gene distal enhancer polymorphism ren-5312c/t is associated with increased blood pressure. Circ Cardiovasc Genet *3*(1), 53–9.

[Vavouri *et al.*, 2006] Vavouri, T., McEwen, G. K., Woolfe, A., Gilks, W. R., & Elgar, G. (2006). Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends Genet. *22*(1), 5–10.

[Venter *et al.*, 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian,

G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T.,

Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., & Zhu, X. (2001). The sequence of the human genome. Science *291*(5507), 1304–51.

[Vlieghe *et al.*, 2005] Vlieghe, D., Sandelin, A., Bleser, P. J. D., Vleminckx, K., Wasserman, W. W., van Roy, F., & Lenhard, B. (2005). A new generation of jaspar, the open-access repository for transcription factor binding site profiles.. Nucleic Acids Res *34*(Database issue), D95–7.

[Warren *et al.*, 2008] Warren, W. C., Hillier, L. W., Graves, J. A. M., Birney, E., Ponting, C. P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A. T., Yang, S.-P., Heger, A., Locke, D. P., Miethke, P., Waters, P. D., Veyrunes, F., Fulton, L., Fulton, B., Graves, T., Wallis, J., Puente, X. S., López-Otín, C., Ordóñez, G. R., Eichler, E. E., Chen, L., Cheng, Z., Deakin, J. E., Alsop, A., Thompson, K., Kirby, P., Papenfuss, A. T., Wakefield, M. J., Olender, T., Lancet, D., Huttley, G. A., Smit, A. F. A., Pask, A., Temple-Smith, P., Batzer, M. A., Walker, J. A., Konkel, M. K., Harris, R. S., Whittington, C. M., Wong, E. S. W., Gemmell, N. J., Buschiazzo, E., Jentzsch, I. M. V., Merkel, A., Schmitz, J., Zemann, A., Churakov, G., Kriegs, J. O., Brosius, J., Murchison, E. P., Sachidanandam, R., Smith, C., Hannon, G. J., Tsend-Ayush, E., McMillan, D., Attenborough, R., Rens, W., Ferguson-Smith, M., Lefèvre, C. M., Sharp, J. A., Nicholas, K. R., Ray, D. A., Kube, M., Reinhardt, R., Pringle, T. H., Taylor, J., Jones, R. C., Nixon, B., Dacheux, J.-L., Niwa, H., Sekita, Y., Huang, X., Stark, A., Kheradpour, P., Kellis, M., Flicek, P., Chen, Y., Webber, C., Hardison, R., Nelson, J., Hallsworth-Pepin, K., Delehaunty, K., Markovic, C., Minx, P., Feng, Y., Kremitzki, C., Mitreva, M., Glasscock, J., Wylie, T., Wohldmann, P., Thiru, P., Nhan, M. N., Pohl, C. S., Smith, S. M., Hou, S., Nefedov, M., de Jong, P. J., Renfree, M. B., Mardis, E. R., & Wilson, R. K. (2008). Genome analysis of the platypus reveals unique signatures of evolution. Nature *453*(7192), 175–83.

[Wasserman & Sandelin, 2004] Wasserman, W. W. & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet *5*(4), 276–87.

[Waterston *et al.*, 2002] Waterston, R., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.,

Brown, D., Brown, S., Bult, C., Burton, J., Butler, J., Campbell, R., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A., Church, D., Clamp, M., Clee, C., Collins, F., Cook, L., Copley, R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K., Deri, J., Dermitzakis, E., Dewey, C., Dickens, N., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D., Eddy, S., Elnitski, L., Emes, R., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G., Flicek, P., Foley, K., Frankel, W., Fulton, L., Fulton, R., Furey, T., Gage, D., Gibbs, R., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T., Green, E., Gregory, S., Guigo, R., Guyer, M., Hardison, R., Haussler, D., Hayashizaki, Y., Hillier, L., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D., Johnson, L., Jones, M., Jones, T., Joy, A., Kamal, M., Karlsson, E., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W., Kirby, A., Kolbe, D., Korf, I., Kucherlapati, R., Kulbokas, E., Kulp, D., Landers, T., Leger, J., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D., Mardis, E., Matthews, L., Mauceli, E., Mayer, J., McCarthy, M., McCombie, W., McLaren, S., McLay, K., McPherson, J., Meldrim, J., Meredith, B., Mesirov, J., Miller, W., Miner, T., Mongin, E., Montgomery, K., Morgan, M., Mott, R., Mullikin, J., Muzny, D., Nash, W., Nelson, J., Nhan, M., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K., Peterson, J., Pevzner, P., Plumb, R., Pohl, C., Poliakov, A., Ponce, T., Ponting, C., Potter, S., Quail, M., Reymond, A., Roe, B., Roskin, K., Rubin, E., Rust, A., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J., Slater, G., Smit, A., Smith, D., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J., Niederhausern, A. V., Wade, C., Wall, M., Weber, R., Weiss, R., Wendl, M., West, A., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R., Winter, E., Worley, K., Wyman, D., Yang, S., Yang, S., Zdobnov, E., Zody, M., & Lander, E. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*(6915), 520–62. 0028-0836 Journal Article.

[Weber *et al.*, 2007] Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., & Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. Nat. Genet. *39*(4), 457–66.

[Wei *et al.*, 2006] Wei, C.-L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T., Shahab, A., Yong, H. C., Fu, Y., Weng, Z., Liu, J., Zhao, X. D., Chew, J.-L., Lee, Y. L., Kuznetsov, V. A., Sung, W.-K., Miller, L. D., Lim, B., Liu, E. T., Yu, Q., Ng, H.-H., & Ruan, Y. (2006). A global map of p53 transcription-factor binding sites in the human genome. Cell *124*(1), 207–19.

[Weiss & Gladstone, 1959] Weiss, S. & Gladstone, L. (1959). A mammalian system for the incorporation of cytidine triphosphate into .... Journal of the American Chemical Society.

[Wells, 2007] Wells, R. D. (2007). Non-b dna conformations, mutagenesis and disease. Trends Biochem Sci *32*(6), 271–8.

[Westerfield *et al.*, 1992] Westerfield, M., Wegner, J., Jegalian, B. G., DeRobertis, E. M., & Püschel, A. W. (1992). Specific activation of mammalian hox promoters in mosaic transgenic zebrafish. Genes Dev *6*(4), 591–8.

[Wickens *et al.*, 1997] Wickens, M., Anderson, P., & Jackson, R. J. (1997). Life and death in the cytoplasm: messages from the 3' end. Curr Opin Genet Dev *7*(2), 220–32.

[Wightman *et al.*, 1993] Wightman, B., Ha, I., & Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in c. elegans. Cell *75*(5), 855–62.

[Woolfe *et al.*, 2005] Woolfe, A., Goodson, M., Goode, D., Snell, P., McEwen, G., Vavouri, T., Smith, S., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y., Cooke, J., & Elgar, G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol *3*(1), e7. 1545-7885 (Electronic) Journal Article.

[Wray *et al.*, 2003] Wray, G., Hahn, M., Abouheif, E., Balhoff, J., Pizer, M., Rockman, M., & Romano, L. (2003). The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol *20*(9), 1377–419. 0737-4038 (Print) Journal Article Review.

[Wray, 2007] Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. Nat Rev Genet *8*(3), 206–16.

[Wu, 1980] Wu, C. (1980). The 5 ends of drosophila heat shock genes in chromatin are hypersensitive to dnase i. nature.com.

[Wu *et al.*, 1979] Wu, C., Wong, Y. C., & Elgin, S. C. (1979). The chromatin structure of specific genes: Ii. disruption of chromatin structure during gene activity. Cell *16*(4), 807–14.

[Yamashita *et al.*, 2005] Yamashita, R., Suzuki, Y., Sugano, S., & Nakai, K. (2005). Genome-wide analysis reveals strong correlation between cpg islands with nearby transcription start sites of genes and their tissue specificity. Gene *350*(2), 129–36.

[Yang & Elnitski, 2008] Yang, M. Q. & Elnitski, L. L. (2008). Prediction-based approaches to characterize bidirectional promoters in the mammalian genome. BMC Genomics *9 Suppl 1*, S2.

[Yu *et al.*, 2006] Yu, X., Lin, J., Zack, D. J., & Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. Nucleic Acids Res *34*(17), 4925–36.

[Yue *et al.*, 2008] Yue, F., Cui, L., dePamphilis, C. W., Moret, B. M. E., & Tang, J. (2008). Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. BMC Genomics *9 Suppl 1*, S25.

[Zhao *et al.*, 2007] Zhao, G., Schriefer, L. A., & Stormo, G. D. (2007). Identification of muscle-specific regulatory modules in caenorhabditis elegans.. Genome Res *17*(3), 348–57.

[Zon, 1999] Zon, L. I. (1999). Zebrafish: a new model for human disease. Genome Res *9*(2), 99–100.