

SECOND RESEARCH DISSERTATION PROJECT



# **Computational Analysis of the Regulatory Region of Vertebrate GSX1**

---

**Candidate Number 675667**

August 2009

**Word count:**

A thesis submitted in partial fulfilment of requirements for the degree of  
MSc Biology (Integrative Biosciences)  
University of Oxford

## ABSTRACT

The computational approach to the study of transcription regulation has become more attractive. It is sufficiently mature to influence the design of the laboratory procedure to detect the transcription factor that regulates the gene and the corresponding transcription factor binding sites. With the accumulation of genome sequences, combined with effective gene-prediction algorithms, comparing genomic sequences across species – phylogenetic footprinting – rapidly become a very effective approach to identify functionally important DNA sequences. Programs such as BigFoot are created based on this principle.

This study employs the computational approach to analyse the regulatory region of vertebrate ParaHox gene, *Gsx1*. Currently, there are very few studies on vertebrate *Gsx1* regulation, and where they do exist, most concentrated on the ParaHox cluster as a whole. Hence, this study is the first to systematically examine the regulatory region of vertebrate *Gsx1*, in particular the regulatory region of human and zebrafish *Gsx1*. With a consensus between three programs (BigFoot, phastCons and ConSite), this study has predicted with confidence five TFBSs in human *GSH1* regulatory region, but none in zebrafish *Gsx1* regulatory region.

# TABLE OF CONTENTS

<b>ABSTRACT</b> .....	<b>2</b>
<b>TABLE OF CONTENTS</b> .....	<b>3</b>
<b>LIST OF FIGURES</b> .....	<b>5</b>
<b>LIST OF TABLES</b> .....	<b>6</b>
<b>ACKNOWLEDGEMENT</b> .....	<b>7</b>
<b>1. INTRODUCTION</b> .....	<b>8</b>
1.1 IDENTIFICATION OF REGULATORY REGIONS .....	9
1.2 COMPARATIVE METHODS .....	12
1.3 ALIGNMENTS AND EVOLUTIONARY MODELS .....	14
1.4 COMBINING STATISTICAL ALIGNMENT AND PHYLOGENETIC FOOTPRINTING .....	19
1.5 METAZOAN GSX1 .....	20
1.6 THE PARAHOX GENE CLUSTER .....	21
1.7 AIMS .....	25
<b>2. MATERIALS AND METHODS</b> .....	<b>26</b>
2.1 DATASET CONSTRUCTION .....	26
2.2 BIGFOOT: STATISTICAL ALIGNMENT AND PHYLOGENETIC FOOTPRINTING SOFTWARE.....	27
2.3 CLUSTAL X SOFTWARE .....	28
2.4 ANALYSIS PROCEDURE .....	29
2.5 COMPARATIVE ASSESSMENT .....	30
<b>3. RESULTS</b> .....	<b>31</b>
3.1 TFBSs PREDICTION: THE COMPUTATIONAL PIPELINE.....	31
3.2 BIGFOOT RESULTS .....	32
3.3 PROMOTER ANALYSIS OF HUMAN AND ZEBRAFISH GSX1 .....	33
3.4 OVERREPRESENTED TF BINDING MOTIFS .....	39
3.5 POTENTIAL NOVEL MOTIFS .....	41
<b>4. DISCUSSION</b> .....	<b>42</b>

4.1 THE PHYLOGENETIC FOOTPRINTING APPROACH.....	42
4.2 STATISTICAL IMPLEMENTATION OF COMPARATIVE APPROACH.....	43
4.3 BIGFOOT: STRENGTH AND LIMITATION .....	44
4.4 REGULATORY ELEMENTS OF VERTEBRATE <i>Gsx1</i> .....	45
4.5 FUTURE WORK .....	46
<b>5. CONCLUSION.....</b>	<b>48</b>
<b>REFERENCES.....</b>	<b>49</b>
<b>APPENDICES .....</b>	<b>52</b>
APPENDIX I : SPECIES INVOLVED .....	52
APPENDIX II : BIGFOOT: ADDITIONAL INFORMATION .....	54
APPENDIX III : PHYLOGENETIC TREES .....	55
APPENDIX IV : BIGFOOT ALIGNMENT RESULTS .....	58

## LIST OF FIGURES

1.1.1	Schematic of a typical gene regulatory region.....	10
1.2.1	Schematic of phylogenetic footprinting.....	13
1.6.1	Metazoan phylogeny with schematics of the genomic organisation of Hox and ParaHox genes.....	22
1.6.2	A model for the maintenance of ParaHox gene cluster.....	23
2.1.1	Regulatory regions analysed.....	27
3.1.1	Strategy for computation based TFBSs prediction analysis.....	31
3.2.1	BigFoot output of two separate analyses of Region 1 (different groups of species).....	32
3.2.2	BigFoot output of two separate analyses of Region 1 (different MCMC parameters).....	32
3.3.1	Regulatory elements prediction for Region 1.....	34
3.3.2	Alignment of human Region 1 with other species by BigFoot.....	35
3.3.3	Regulatory elements prediction for Region 2.....	36
3.3.4	Regulatory elements prediction for Region 3.....	37
3.3.5	Regulatory elements prediction for Region 4.....	37
3.3.6	Regulatory elements prediction for Region 5.....	38

## LIST OF TABLES

3.4.1	TFBSs predicted by ConSite program.....	39
3.4.2	Summary of overrepresented TF binding motifs as predicted by ConSite program.....	40
3.5.1	Potential novel binding sites for the 5 regions.....	41

## **ACKNOWLEDGEMENT**

I would like to express my gratitude to Prof. Jotun Hein, for supervising this project. Thanks for sparing your time to answer all my enquiries and for the invaluable feedbacks.

I am grateful to everyone in the Genome Analysis and Bioinformatics group, especially to Adam Novak and Rune Lyngsoe for making the journey into the world of mathematics and computer science enjoyable (and comprehensible!) for a biologist like me. I have learnt a lot from this group despite the short time period.

I also appreciate the helpful discussions with John Mulley and Prof. Peter Holland who are working on the ParaHox cluster.

I would also like to thank Nanette Coetzer for the tips and discussions on the project, and her group mates, Cathleen Heil, Gabor Boross and Andras Gyorgy, for sharing their report.

I gratefully acknowledge those whose work, thesis and reports had helped me in the making of my report.

Finally, I would like to thank my family and friends for their moral support.

# Computational Analysis of the Regulatory Region of Vertebrate GSX1

## 1. INTRODUCTION

The application of information technology to the field of molecular biology has revolutionized our approach to the study of genetics. Over the past three decades, we are witnessing the fast advancement of bioinformatics techniques, inspired by success stories of multidisciplinary approaches in problem-solving. Indeed, rapid accumulation of complete multiple metazoan genome sequences, overwhelmed by large-scale expression data, have combined to motivate researchers in developing better bioinformatics methods, such as for the analysis of transcriptional regulation networks.

The concept of gene itself has changed a great deal over time. In classical genetics, a gene is a unit of inheritance that transmit a characteristic from parent to child. Then came the field of biochemistry, which associates the transmitted characteristic with proteins, usually one for each gene. With the advent of molecular biology, genes became sequences of DNA that are commonly seen as the functional portion in the genome, which are transcribed into messenger RNA (mRNA) and eventually translated into protein. Of course, this is a simplified – and highly publicized – picture of the gene model. As a result, less attention has been given to the non-protein-coding regions and elements of the genome which temporally and spatially regulate gene expression (Greally, 2007). However, we eventually came to realise that, in order to understand the human genome and by extension the biological processes it orchestrates, a more transparent view of the information encoded in the genome is required.

When the draft human genome sequence was completed in 2001, the world of genetics was shaken by several of the surprising outcomes. For instance, the number of protein-coding genes was much lower than expected, comprising only about 1.5% of our genome (Birney *et al.*, 2007; Greally, 2007; Mattick, 2003). Additionally, the fact that 95% of human protein-coding genes are similar to those found in mouse suggests the presence of conservation of protein-coding genes between different organisms (Mattick, 2003). However, one important finding from the project has forced us to start paying attention to the non-protein-coding regions – that variations, whether in humans or in other organisms, occur not only in the protein-coding region of the genome, but also the non-protein-coding region.

Then geneticists eventually discover that most genetic diseases are more likely to be due to genetic variation in regions that control activity of genes, rather than in the regions that specify the protein code (Dimas, 2009). Although polymorphism in protein-coding sequences may alter the structure and function of the protein, it is the difference in the regulatory architecture that not only affects the patterns and amounts of protein expressed in different cells and tissues, but to some extent, influence our physical and physiological characteristics (Mattick, 2003). Thus, if we are to understand the insight into the causes of human diseases, knowing the functions and attributes of the non-protein-coding region become necessary.

There is also a growing body of evidence showing that the gene regulatory systems underlie the evolution of morphological diversity more than changes in gene number or protein function (Carroll, 2000). Hence, unsurprisingly, one of the most widely studied processes in cell and molecular biology is the initial step of gene expression – transcription.

## **1.1 Identification of regulatory regions**

Whether using the traditional, empirical approaches or mature bioinformatics methods, the principle being exploited remains similar. The major task of deciphering a transcriptional regulatory system is to identify all transcription factor binding sites (TFBSs) bound by the transcription factors (TFs) encoded in the genome (Figure 1.1.1). It is important, however, to keep in mind that characterizing the mechanisms that govern initiation of transcription does not reveal the entire picture, since regulation of any genes may occur after transcription of pre-mRNA (for instance, through splicing and protein modification).

Most of the conventional molecular techniques have always depended on the idea that TFBSs bound to the TFs will have different properties (i.e. molecular weight, disintegration rate when exposed to certain chemicals, etc.) (Cohen *et al.*, 1986; Lin *et al.*, 2007), compared to other non-functional sequences and unrelated TFBSs. Additionally, regulatory regions have been empirically predicted on the basis of chromatin structure, since several studies have demonstrated the strong association between transcription start sites (TSSs) and both histone modifications and DNAase hypersensitive sites (Birney *et al.*, 2007). In general, researchers would have to incorporate several experiments to be able to identify the regulatory regions of genes analysed.

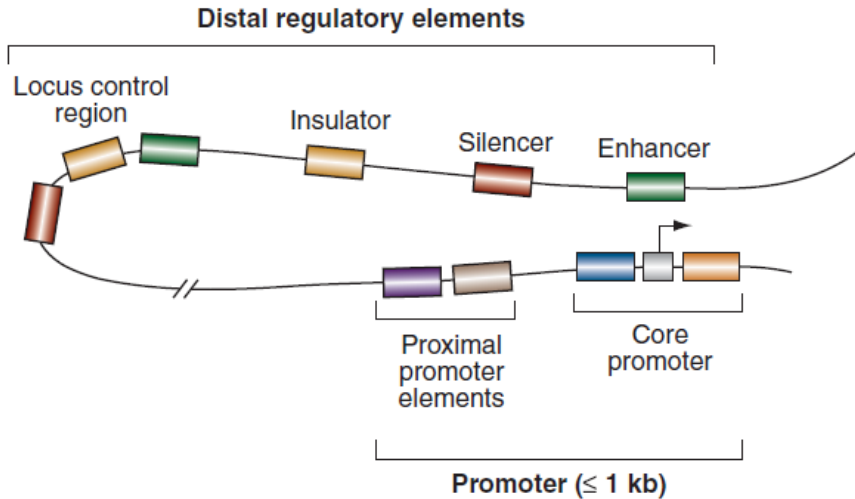


Figure 1.1.1 Schematic of a typical gene regulatory region. In eukaryotes, TFs bind to specific sites (TFBSs) that are either proximal or distal to a transcription start site, and mediate the binding of RNA polymerase, forming the transcription initiation complex. The promoter, which is composed of a core promoter and proximal promoter elements, typically spans less than 1000 bp. Distal regulatory elements which can include enhancers, silencers, insulators, and locus control regions, can be located up to 1 Mb pairs from the promoter. (Image from Maston *et al*, 2006).

It is these laborious laboratory procedures that have made the computational approach more attractive. Furthermore, the bioinformatics methods that address the initiation of transcription are sufficiently mature to influence the design of the laboratory procedures.

The initial step towards the identification of TFBSs is to distinguish larger regions that might harbour regulatory elements, typically upstream the 5' end of any protein-coding gene, and find the promoter – the upstream region proximal to a gene's TSS (Qiu, 2003). The promoter detection algorithms, in general, can be classified into two groups. First is the signal-based approach, which relies on the recognition of relatively conserved signals and conserved spacing among patterns. Generally, underlying most algorithms for this promoter prediction process is a reference collection known as the 'Eukaryotic Promoter Database' (EPD) (Wasserman and Sandelin, 2004). Signals frequently used for this kind of detection include TATA box sequences, which are often located about 30 base pairs (bp) upstream a TSS; and CCAAT box sequences, which signals the binding site for the RNA transcription factor and located about 75-80 bp upstream a TSS. An alternative approach, tested in several programmes such as ProScan (<http://www-bimas.cit.nih.gov/molbio/proscan/>), is using the weight matrix or nucleotide distribution matrix. Weight matrices are selective description of

DNA patterns, where each nucleotide position are accurately weighed according to the observed biological conservation (Cartharius *et al.*, 2005).

The second group of algorithm utilize the content-based approach. Promoter sequences are distinguished from non-promoter sequences based on content differences such as triplet base-pair preferences around TSS or the hexamer frequencies in consecutive 100-bp upstream regions, by using statistical classification methods such as linear discriminant function or quadratic discriminant analysis (Qiu, 2003; Zhang, 1998). Usually, a higher-resolution result can be achieved this way. For example, when given a 1- to 2-kb extended promoter, the program CorePromoter (<http://rulai.cshl.org/tools/genefinder/CPROMOTER/index.htm>) will be able to correctly localize TSS to a 100-bp interval 60% of the time (Qiu 2003).

Despite the moderate accuracy in prediction (about 13 – 60%), most of these early programs are plagued by false positives. Hence, the next generation of algorithms has shifted the emphasis on predicting promoters, by developing models based on context features extracted from computational machine learning algorithms (Box 1) (Wasserman and Sandelin, 2004). Additional information is also combined in these algorithms, as empirical data on dominant characteristic of promoter sequences in the human genome accumulate. One such data is the abundance of CpG dinucleotides, since many vertebrates promoter regions coincide with CpG islands. These short, unmethylated, GC-rich, CpG-dense regions of DNA usually located at the 5' ends of genes, potentially, mark the position of the large proportion of promoters in mammalian genomes (Takai and Jones, 2002). Computationally, this imbalance of CG dinucleotides can be a powerful tool for finding regions in genes that are likely to contain promoters (Wasserman and Sandelin, 2004).

The strategy of identifying regulatory regions via promoters, however, does come with a drawback. For one thing, promoters are very diverse and even well-known motifs are not always conserved in all promoters (Qiu 2003). Additionally, when (Venter *et al.*, 2001) published their version of the human genome, they found that the human genome contains about 1850 unique TFs, each of which binds to specific TFBSs. Since each gene also contains a set of unique combination of TFBSs in the promoter to determine its temporal and spatial expression, the number of combinatorial pattern permutations can be immense.

Therefore, it is natural that the next step to do after promoter regions have been identified is to search for *cis*-regulatory elements computationally. One common approach is to use

position weight matrix (PWM) to quantitatively represent binding specificity, since most TFs bind to short (6 – 25 bp) degenerate sequence motifs. Several programs developed to perform searches on an input sequence based on the PWM include MatInspector, SIGNAL SCAN and TFSearch (Qiu 2003).

Indeed, as the empirical technology in the analysis of transcriptional regulatory region becomes more advanced, so do the algorithms, to cope with the increasing datasets. For instance, researchers eventually noticed that gene can be classified into different clusters based on their expression patterns, hence those belonging to the same cluster are assumed to be co-regulated. In other words, supposed that co-regulated genes are under similar regulatory control, then algorithms can be polished to search for overrepresented motifs in the collection of regulatory regions (Roth *et al.*, 1998). Popular programs known to be able to perform this task include MEME, Gibbs Sampler, ANN-Spec, etc. (Qiu, 2003). Keep in mind, however, that not all co-regulated gene promoters share common motif – some of the genes in a given cluster might be secondary response genes. Additionally, same motif might be found in the regulatory regions of genes of other clusters, since this approach is also subjected to limitations of expression profiling experiments.

The methods discussed thus far only work best when identifying likely TFBSs in one species, because the approaches are based on the alignment of regulatory regions of co-regulated genes, or the statistical analysis of overrepresented motifs. Application on multi-species will be more complex, hence resulting in a higher frequency of false positives (Qui 2003).

## **1.2 Comparative methods**

To address the aforementioned complexities of identification of TFBSs using more than one species for analysis, researchers came up with comparative methods – also referred to as phylogenetic footprinting – which, in fact, is an approach inspired by the wet-lab technique of DNase footprinting (Zhang and Gerstein, 2003). As data on genomes of several different species built up, it has become obvious that cross-species genome comparison is a very effective way to identify functionally important DNA regions.

This approach relies on the assumption that regulatory elements, just like protein-coding sequences, are under evolutionary selection, which means these functionally important regions should have evolved much more slowly than other non-coding sequences (Qiu, 2003;

Wasserman and Sandelin, 2004; Zhang and Gerstein, 2003). Orthologous genes, therefore, will be subject to the same regulatory mechanisms in different species. This hypothesis has been validated by several studies of individual genes and genome-wide sequence comparison (e.g. (Lenhard *et al.*, 2003; Prakash and Tompa, 2005), and can be illustrated schematically as shown in Figure 1.2.1.

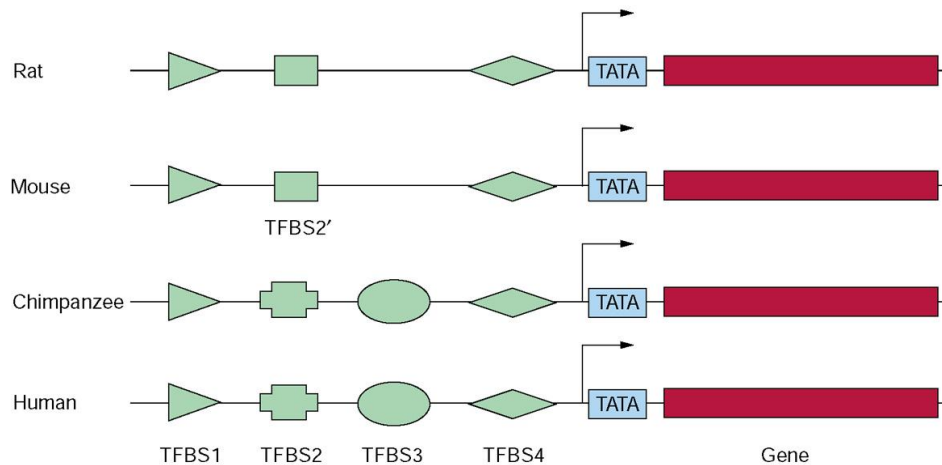


Figure 1.2.1 Phylogenetic footprinting. This schematic diagram shows a hypothetical human gene aligned with its orthologs from chimpanzee, mouse and rat. Sequence motifs of the same shape (green) represent binding sites of the same class of TFs. TFBS1 and TFBS4 are conserved in all mammals; TFBS3 represent newly acquired, primate-specific binding site; and TFBS2 and TFBS2' represent orthologous regulatory sites that have diverged significantly between primate and rodent lineages. Conserved TFBSs in each sequence are revealed through cross-species sequence comparison. (Image from (Zhang and Gerstein, 2003))

In the beginning, phylogenetic footprinting was performed by visually examining the alignment of orthologous sequences. Then, automated computer programs were developed to assist the process. In 2004, (Sandelin *et al.*, 2004) introduced an interactive web-based computational platform called ConSite (<http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite/>), which allows users to do their own phylogenetic footprinting. The phylogenetic footprinting approach has been well-accepted in the research community because a large number of false positives, such as those produced by the PWM approach, are reduced. Hence, it is common for this phylogenetic filtering procedure to be used together with other approaches. On average, phylogenetic footprinting improved the selectivity of TFBS prediction by 85% compared to matrix models alone (Zhang and Gerstein, 2003).

The major advantage of phylogenetic footprinting is its capability to identify regulatory elements specific even to a single gene; while the single-genome, multigene approach, on the other hand, requires a reliable set of co-regulated genes. It is important, however, to be aware

of the evolutionary distances between the different species being investigated. In general, the alignment-based phylogenetic footprinting approach works best for orthologous genes from species with appropriate evolutionary divergence. Too closely related species, such as human-chimpanzee, provides little benefit since the sequences closely resemble each other; but too widely diverged species, such as primate-fish, can show no detectable similarity (Wasserman and Sandelin, 2004). Indeed, since little was known about the relative merits of evolutionarily close and distant sequence comparisons, Prabhakar *et al* (2006) used a uniform computational approach (Gumby) to investigate the impact of evolutionary distance on predictive power, and compared the outcomes against previously identified regulatory elements. Their results highlighted the practical utility of close sequence comparisons, and the loss of sensitivity entailed by more distance comparisons (Prabhakar *et al.*, 2006).

Many current research directions are currently being pursued in this area of phylogenetic footprinting. For instance, researchers are now interested in expanding pair-wise sequence comparisons to include multiple species, and to make use of additional information, such as evolutionary distance (Blanchette *et al.*, 2002), structural information (Wasserman and Sandelin, 2004) and functional information (Cooper and Brown, 2008).

### **1.3 Alignments and evolutionary models**

In general, the phylogenetic footprinting algorithm is made of three components; defining the suitable orthologous gene sequence for comparison, aligning the promoter sequences of orthologous genes, and identifying segments of significant conservation (Wasserman and Sandelin, 2004).

Sequence alignment generally fall into two categories; one that targets short segments of similarity (local alignment), while the other aims for an optimal description of similarity across the entire pair of sequences (global alignment) (Wasserman and Sandelin, 2004). Other alignment methods include pairwise alignment, used to find the best-matching local or global alignments of two query sequences, and multiple sequence alignment, which is an extension of pairwise alignment to incorporate more than two sequences at a time. A variety of computational algorithms have been applied to the sequence alignment problem and have been implemented in programs such as BLAST and PIPMaker which are based on local alignments, and programs such as CLUSTALW and Vista which are based on global alignment (Qiu 2003).

The power of phylogenetic footprinting depends on our ability to align the orthologous region of interest obtained in order to identify the segments of similarity, and to construct and interpret phylogenetic trees. Statistical methods are applied to determine the likelihood of a particular alignment between sequences arising by chance given the size and composition of the database being searched.

To produce good alignments, researchers have proposed several models to reflect the process of how the sequence came about. Commonly, Markov chains (or Markov processes) are used to model the evolution of biological processes such as population dynamics, evolution of quantitative traits and evolution of genomic data (Galtier *et al.*, 2005). Having a Markovian property means that future states depend only on the present state, and are independent of the past states. Since Markov chains are stochastic processes, change of states (transitions) are probabilistic. In Markov models of DNA evolution, transitions between one nucleotide to the other are defined by its rate matrix. These models, therefore, mostly differ in the parametrization of the rate matrix and in the modelling of rate variation.

The first – and simplest – model of nucleotide evolution is the (Jukes and Cantor, 1969) (JC69) model, which assumes that the rate of substitution is the same between all nucleotides. Therefore, the model only requires a single parameter-denoting rate,  $\mu$ , which is the overall substitution rate, and a value for time. The rate matrix (Q) will be as follow;

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

Where the transition probability matrix (P(t)) when  $t > 0$  will be as follow;

$$P(t) = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

However, this is a strong assumption that turned out to fit virtually no sequence dataset. Kimura (1981) amended the JC69 model to incorporate two model parameters; one for the

transition (purine to purine, or pyrimidine to pyrimidine ( $\kappa$ )) and one for the transversion rate (purine to pyrimidine or vice versa (1)) (K80). The rate matrix (Q) will now be as follow;

$$Q = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

The JC69 and K80 models both have a balanced stationary distribution (i.e. at equilibrium, all the four bases are expected to have a proportion of 0.25 each). Many DNA sequences dataset, however, show unbalanced base composition, so Felsenstein (1981) introduced F81 model, in which the substitution rate corresponds to the equilibrium frequency of the target nucleotide, which assumes unequal base frequencies ( $\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$ ). The rate matrix (Q) is as follow;

$$Q = \begin{pmatrix} * & \pi_T & \pi_T & \pi_T \\ \pi_C & * & \pi_C & \pi_C \\ \pi_A & \pi_A & * & \pi_A \\ \pi_G & \pi_G & \pi_G & * \end{pmatrix}$$

Hasagawa *et al* (1985) eventually proposed a unified version of the last two models, producing a five parameter model (HKY85), with a rate matrix (Q) as follow;

$$Q = \begin{pmatrix} * & \kappa\pi_T & \pi_T & \pi_T \\ \kappa\pi_C & * & \pi_C & \pi_C \\ \pi_A & \pi_A & * & \kappa\pi_A \\ \pi_G & \pi_G & \kappa\pi_G & * \end{pmatrix}$$

By 1990s, most of the models used are the developed and refined version of the HKY85 model, such as the TN93 model proposed by Tamura and Nei in 1993, which included parameters that would distinguish between the rates of pyrimidines and purines respectively.

Despite the development in models of nucleotides evolution, none of the early alignments algorithms incorporated the stochasticity of real evolutionary processes, or reached the general agreement on the assessment of the significance of the alignments achieved (Bishop and Thompson, 1986). In fact, statistical approach in the alignment of sequences has been deemed to be computationally too slow in the early 1990s (Hein *et al.*, 2000), so much of the original bioinformatics techniques were based on global multiple alignment, produced using

the Needleman-Wunsch algorithm (Box 1) which assumes that important functional sequences will remain collinear over evolution (i.e. in the same order and orientation along the gene) (Wasserman and Sandelin, 2004).

### **Box 1. Glossary**

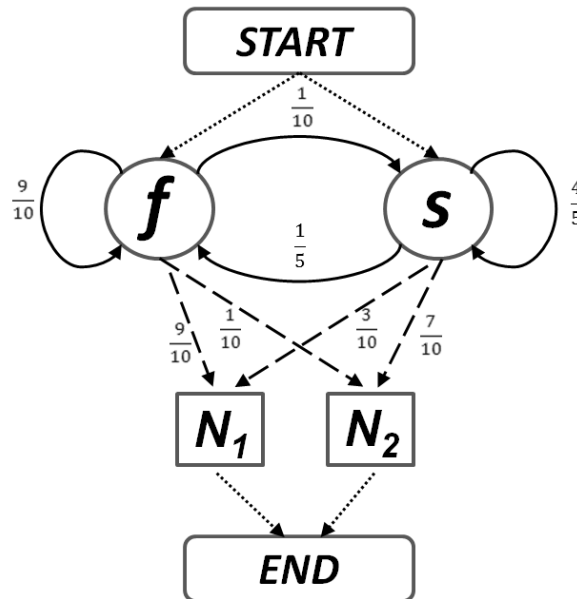
- ***Computational machine learning algorithm***  
Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow programs to learn from experience – that is, to modify its execution on the basis of newly acquired information. In bioinformatics, neural networks and MCMC are well-known examples.
- ***Needleman-Wunsch algorithm***  
The Needle-Wunsch algorithm performs a global alignment on two sequences. It is commonly used in bioinformatics to align proteins or nucleotide sequences. The algorithm returns an optimal alignment, in which ‘optimal’ refers to the highest possible score under a specific scoring system. The algorithm is an example of dynamic programming, and is computationally demanding, hence restricting its direct application to sequences of modest length.
- ***MCMC sampling***  
MCMC methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. After several steps, the state of the chain is used as sample for the desired distribution. Common problem with MCMC methods, then, is to determine how many steps are needed to converge to the stationary equilibrium within an acceptable error.

In 1986, Bishop and Thompson proposed the first statistical alignment procedure, which used a maximum likelihood solution to the DNA sequence alignment problem under a stochastic model. Using a similar approach, Thorne *et al* introduced the TKF91 model in 1991, which is a continuous-time evolutionary model for sequence insertions, deletions and substitutions. According to this model, there will not be one alignment, but all possible alignments, to contribute to the likelihood of the two observed sequences (Thorne *et al.*, 1991). A major advantage of the model is that it can be treated analytically. Indeed, the model can be reformulated as a hidden Markov model (HMM) (Box 2), a statistical model in which the system being modelled is assumed to be a Markov process with unobserved state (Rabiner, 1989).

## Box 2. Hidden Markov Model

A process is said to be Markovian if the next state only depends on the current state, and indeed, in bioinformatics a lot of phenomenon are perceived to be, or at least modelled as, Markovian in nature. For instance, in nucleotide substitution models, the rate of change to a particular nucleotide may depend on the current nucleotide but not on the nucleotides previously occupying the position.

In a hidden Markov model, a sequence of observations (e.g. denoted  $N$ ) is available to be analyzed, but the sequences of states (e.g. denoted  $f$  and  $s$ ) by which the observations were generated is 'hidden'. An example of a HMM for a DNA sequence containing two types of nucleotides (for simplicity) is as follow;



Hence, a run following a sequence of states  $ffs sf$  would emit a sequence of observations  $N1 N1 N2 N1 N1$ . Special Start- and End- states are added to generate finite outputs. The transition probabilities are required to specify a probability distribution over outgoing transitions for each state. Formally, the HMM  $M$  is defined with a tuple  $(Q, \Sigma, a, e, \text{start}, \text{end})$  where

- $Q$  is the set of states in the model (e.g.  $f$  and  $s$ )
- $\Sigma$  is the set of possible observations (e.g.  $N1$  and  $N2$ )
- $a_{p,q}$  describes the transition probability between states  $p$  and  $q$  in  $Q$  (e.g.  $a_{f,s} = \frac{1}{5}$ )
- $e_{p,\sigma}$  for  $p \in Q$  and  $\sigma \in \Sigma$  describes the emission probability of symbol  $\sigma$  when entering state  $p$  (e.g.  $e_{s,N1} = \frac{3}{10}$ )
- $\text{start}, \text{end} \in Q$  are designated start and end states of the model.

Efficient algorithms are available to compute the maximum-likelihood path, the posterior probability that any given state generated any given element of  $N$ , and the total probability of  $N$  considering all possible states (i.e. the likelihood of the model).

## 1.4 Combining statistical alignment and phylogenetic footprinting

Researchers now realised that incorporation of statistical alignment to phylogenetic analysis can significantly improve the accuracy of detecting conserved regions. One such program to do so is phastCons (Siepel *et al.*, 2005). The basis of this program is modelled by phylogenetic HMM, which considers the phylogeny of the species being aligned in their HMM. Basically, the program scans along the alignment for conserved regions, based on three factors; the multiple alignment, the phylogenetic model of conserved regions, and the phylogenetic model for non-conserved regions.

Then a step further, Satija *et al.* (2008) created SAPF, a new program for annotating regulatory elements, by combining a statistical aligner and phylogenetic footprinter. SAPF has the ability to (i) distinguish between sequence undergoing neutral evolution and conserved sequences that may be the result of purifying selection, (ii) calculate and sum up a distribution of many possible alignments, (iii) analyze sequence data from multiple organisms, related by any previously known phylogenetic tree, and (iv) model insertion and deletion events of multiple nucleotide fragments with geometrically distributed lengths, with a distribution parameter that can be set in the model (Satija *et al.*, 2008). Tests on SAPF have demonstrated the potential of combining statistical alignment with phylogenetic footprinting to improve the accuracy of regulatory signal detection.

However, due to the large number of states in the SAPF HMM, the program can only analyze data from four different species, hence placing SAPF at a disadvantage to other methods that can analyze data from significantly greater numbers of species, such as phastCons. One suggested improvement is to use Markov Chain Monte Carlo (MCMC) (Box 1) simulation techniques to approximate the alignment probability distribution (Satija *et al.*, 2008). This novel approach was implemented in a software package BigFoot (see Methods), recently released by the Genome Analysis and Bioinformatics Group, University of Oxford (Satija *et al.*, in press). Preliminary analysis has demonstrated that BigFoot has the potential to outperform existing alignment-based phylogenetic footprinting techniques, and can be applied to a variety of biological databases (Satija *et al.*, in press).

Once an alignment or set of alignments is defined, the next step is to interpret the data obtained, with output usually delivered through graphical display. In general, if none or not all of the well-conserved regulatory motifs have significant matches to known TFBSs, such

as those stored in TRANSFAC (experimentally verified TFs database), then they may be excellent candidates for novel functional regulatory elements.

## 1.5 Metazoan GSX1

Gsx is a class of homeobox genes, found in a diversity of proteosomes and deuterosomes, which implies that the gene class date at least to the base of the Bilateria (Ferrier and Minguillon, 2003). Similar to Nkx and Msx class homeobox genes, Gsx genes are components of a conserved regulatory network involved in patterning the columnar organization of neuronal precursors in both protostome and deuterostome lineages (Illes *et al.*, 2009).

Duplication during the evolution of vertebrates from primitive chordates resulted in two Gsx genes to be found in amniotes (*Gsh-1* and *Gsh-2*). The human *Gsx-1* maps to chromosome 13, at position 13q12.1 (Mulley *et al.*, 2006). Mutant phenotype in several vertebrates indicates that Gsh-1 gene is critical in the genetic hierarchy of the formation and function of the hypothalamic-pituitary axis. Analysis of *Gsh-1* knock out mice shows that Gsx function is required for the proliferation of subsets of neuronal precursors in the ventricular zone of the ventral telencephalon (cerebral cortex). The target genes of *Gsh-1* include the growth hormone-releasing hormone (GHRH) gene (Li *et al.*, 1996), and achaete-scute complex-like 1 (*Ascl1*) gene (Wang *et al.*, 2009).

Expression analysis of *Gsh1* in *Xenopus tropicalis* revealed patterns of expression within the anterior nervous system (Illes *et al.*, 2009). In fact, previously reported expression patterns of chordate Gsx class genes shows that expression in the most anterior regions of the nervous system is a common feature (Illes *et al.*, 2009). In amphioxus, Gsx gene is expressed in a single domain, the cerebral vesicle, which is the anatomical homologue of the vertebrate forebrain/midbrain (Brooke *et al.*, 1998).

Phylogenetic analysis indicates that Gsx genes are mostly closely related to anterior Hox genes of paralogue groups 1 and 2 (Brooke *et al.*, 1998). Hence, the Gsx gene was originally labelled an orphaned Hox-gene due to its location being dispersed away from other known Hox clusters. However, in 1998, Brooke *et al* revealed that the Gsx genes in amphioxus (*Branchiostoma floridae*) were actually located in close proximity with the other two classes

of homeobox genes (i.e. Pdx and Cdx), hence forming a tight cluster which they labelled as the ParaHox gene cluster.

## 1.6 The ParaHox gene cluster

The presence of ParaHox gene cluster challenges the idea that Gsx, Xlox and Cdx gene classes are ‘dispersed’ Hox genes. Molecular phylogenetic analyses of the three gene classes demonstrated that the ParaHox gene cluster is an ancient paralogue of the Hox gene cluster – that the two gene clusters arose by duplication of a ProtoHox gene cluster (Brookes *et al.*, 1998). Indeed, it has been shown in amphioxus that the ParaHox cluster also exhibits both Spatial and Temporal Colinearity, although the Temporal Colinearity is inverted with respect to the pattern in the Hox cluster (Ferrier and Minguillon, 2003).

So far, the information on the ParaHox cluster is restricted to the chordates. Conventional model systems (i.e. *Drosophila melanogaster* and *Caenorhabditis elegans*) are not helpful in understanding ancestral features of the ParaHox gene cluster because they have lost the cluster and deleted some of the genes (Ruvkun and Hobert, 1998) (Figure 1.6.1). This cluster disintegration situation is reflected in the Hox cluster of the two animals as well (Ferrier and Minguillon, 2003).

Homeobox gene families in an amphioxus are not complicated by either excessive duplication or divergence and rearrangement, so just like its Hox cluster, amphioxus has a single ParaHox gene cluster only, containing all the three genes (Brookes *et al.*, 1998). In mammals (i.e. human and mouse), on the other hand, there are two Gsx genes (*Gsh-1* and *Gsh-2*), a single Pdx class genes (*Xlox*), and three Cdx genes (*Cdx1*, *Cdx2*, and *Cdx4*) in the genome. However, only a single intact ParaHox cluster contained all the three classes of homeoboxes, while the remaining genes exist in degerate clusters containing single ParaHox genes (Ferrier *et al.*, 2005). This resulted from ParaHox gene cluster duplications at the base of the vertebrate lineage, and presence of the ParaHox gene cluster has been empirically confirmed in amphioxus, humans, mouse and *Xenopus* (Mulley *et al.*, 2006).

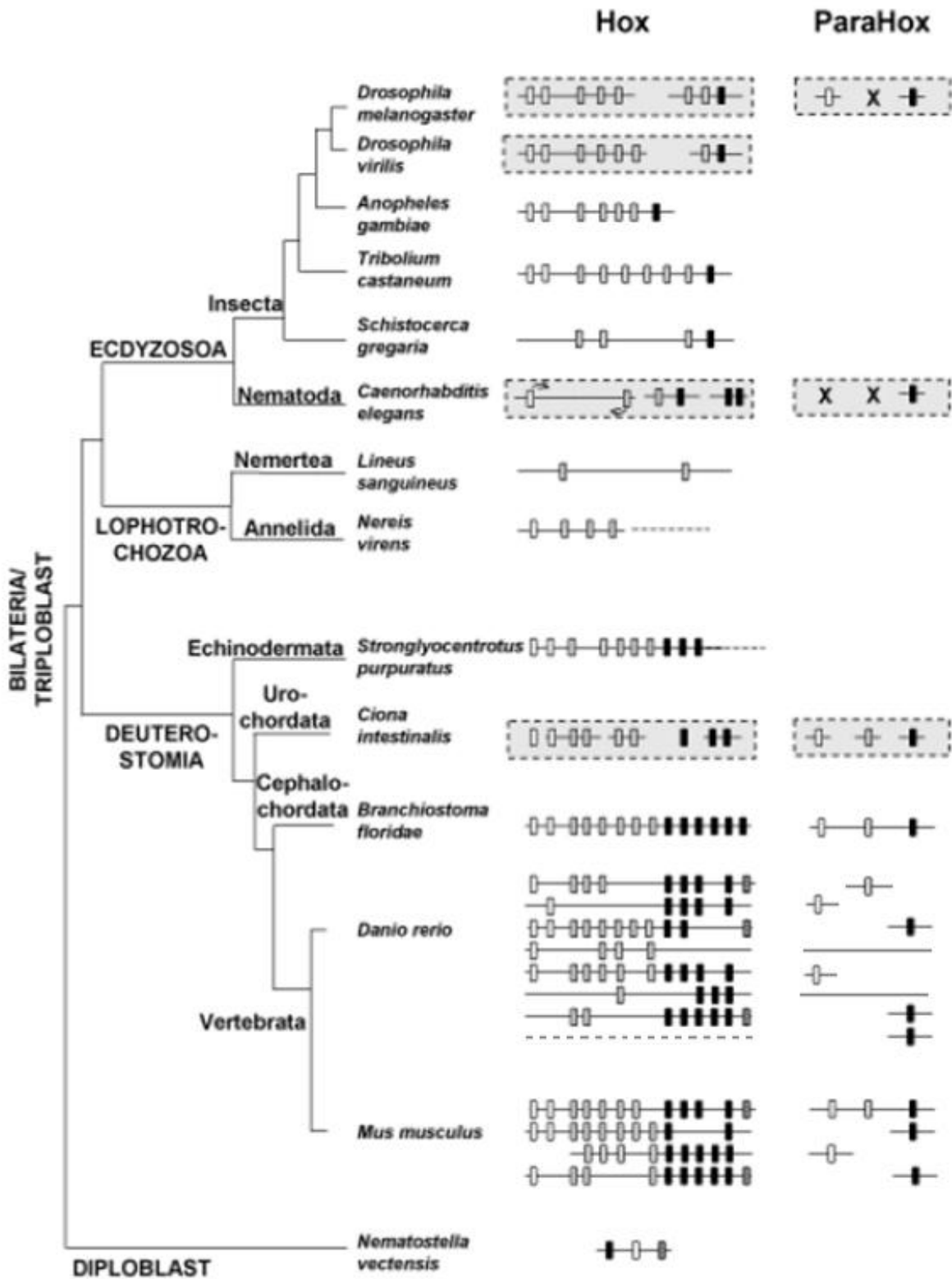


Figure 1.6.1 Metazoan phylogeny with schematics of the genomic organisation of Hox and ParaHox genes. The Hox and ParaHox clusters that have broken up are highlighted in grey boxes. White genes – anterior Hox and Gsx; grey genes – group 3 central and Xlox/Pdx; black genes – posterior Hox and Cdx. X represents loss of a ParaHox gene. Whole-genome duplication in the fish actinopterygian lineage resulted in four additional cluster, but one was loss in *D. rerio* (dashed line). (Adapted from Ferrier and Minguillon, 2003)

It has been implied that there is a strong selective pressure to maintain the physical linkage of the three homeobox genes in at least one copy of the ParaHox clusters; otherwise, inversions and translocations would have dispersed the genes during the half-billion years since the divergence of cephalochordates and vertebrates (Mulley *et al.*, 2006).

However, Mulley *et al* (2006) showed that in the ray-finned fish clade, the ParaHox gene cluster was lost in the evolution of teleosts after divergence from more basal ray-finned fish. Hence, Mulley *et al* (2006) proposed that the ParaHox gene cluster is held together in chordates by the existence of interdigitated regulatory regions that could be separated after locus duplication in the teleosts fish, which is tolerated because of genetic redundancy of both genes and regulatory elements (Figure 1.6.2).

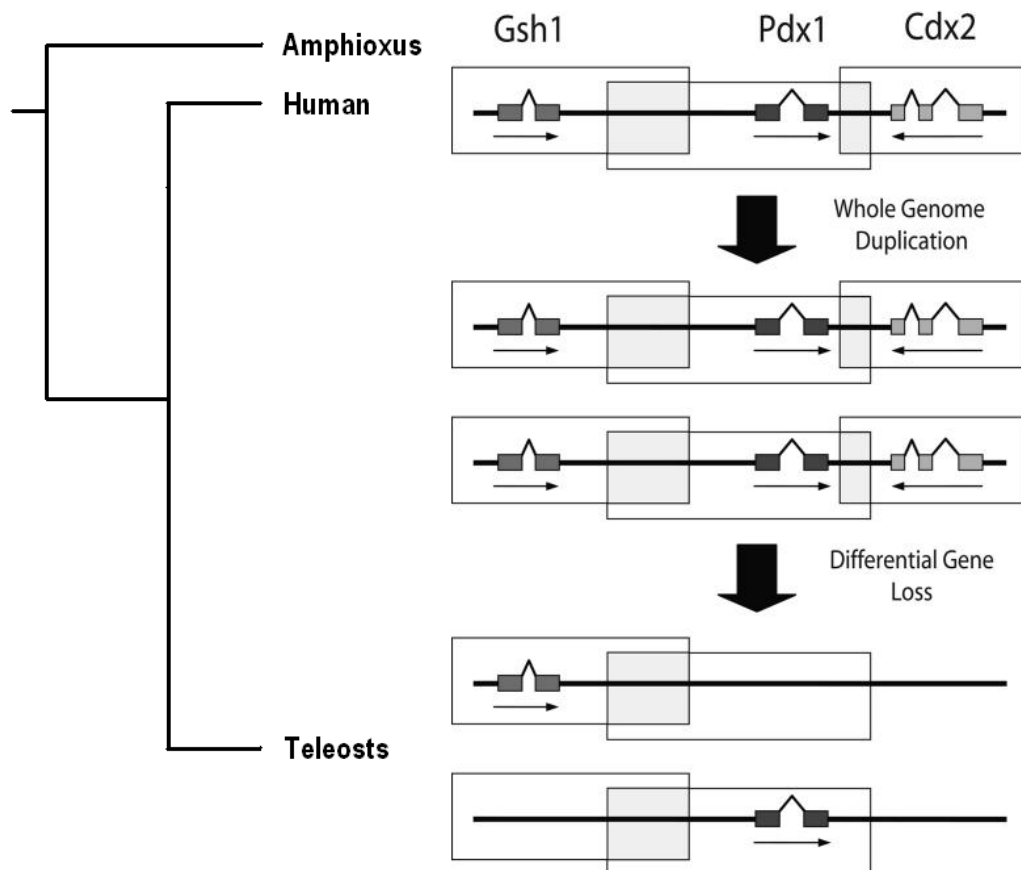


Figure 1.6.2. (Right) A model for the maintenance of ParaHox gene clustering, based on interdigitated and/or shared enhancers. The clear rectangles represent the proposed regulatory regions for each gene. (Left) Evolution of the vertebrate ParaHox cluster. From a single cluster of three genes in amphioxus, an intact cluster of ParaHox genes is conserved between amphibians and mammals. The ParaHox cluster is broken in teleosts after whole-genome duplication. (Adapted from Mulley *et al*, 2006)

Consistent with their predicted model, Mulley *et al* (2006) also detected a large conserved non-coding region sited between *Gsh1* and *Pdx1*, although experimental tests are needed to determine its function (Mulley *et al.*, 2006). Currently, the Evolution and Development Research Group, University of Oxford, is experimentally working on the functional roles of the highly conserved non-coding sequences in the ParaHox cluster of vertebrates and how whole-genome duplication might affect the organization and expression of these clustered genes. To date, no TFBSs for the promoter region of either human or zebrafish *Gsx1* have been verified experimentally by the group. However, according to the BioBase database [\(link?\)](#), there is only one known TF so far that is involved in the regulation of this gene in [\(spp?\)](#), namely the transcription factor SP1. The TFBS for SP1 is located in [\(where?\)](#) and has been detected using chromatin immunoprecipitation (ChIP) assay. Therefore, the lack of a priori on *Gsx1* regulation, together with the progressively more sophisticated computational approach in promoter analysis means that we can get an integrated platform for deciphering the transcriptional regulatory elements of the vertebrate *Gsx1* gene.

## 1.7 Aims

This project aims to consider the use of computational approach in the analysis of the promoter region of human and zebrafish Gsx1 gene. The specific objectives are as follow:

- (i) To detect conserved regions, which can potentially be a TFBS, on the non-coding region adjacent to Gsx1 gene for human and zebrafish
- (ii) To test the reliability of BigFoot as a sophisticated program in regulatory element detection
- (iii) To compare BigFoot analysis results with phastCons, another well-known TFBSs detection program
- (iv) To investigate the reliability of comparative methods in detecting conserved regions
- (v) To utilize online Bioinformatics tools to collect relevant data for analysis.

## 2. MATERIALS AND METHODS

### 2.1 Dataset construction

The non-coding DNA regions analysed in this project were based on five main segments (Figure 2.1.1),

- (i) **Region 1** – Human (hg18) Chr13: 27,263,780 - 27,264,780  
1000 bps upstream of the start site of mRNA transcript of the human GSH1 gene
- (ii) **Region 2** – Human (hg18) Chr13: 27,266,089 - 27,267,089  
1000 bps downstream of the stop site of mRNA transcript of the human GSH1 gene
- (iii) **Region 3** – Human (hg18) Chr13: 27,259,902 – 27,260,253  
a region of about 350 bp long, located about 1 Mbp upstream of human GSH1 gene where known TFBS has been detected empirically by ChIP assay
- (iv) **Region 4** – Zebrafish (danRer4) Chr5: 75,391,548 - 75,392,548  
1000 bps upstream of the start site of mRNA transcript of the zebrafish Gsx1 gene
- (v) **Region 5** – Zebrafish (danRer4) Chr5:75,389,094 - 75,390,094  
1000 bps downstream of the stop site of mRNA transcript of the zebrafish Gsx1 gene

The sequences were acquired from the University of California Santa Cruz (UCSC) Genome Browser website (<http://genome.ucsc.edu/index.html>). For each species, sequences of homologous region in other species included in the Conservation (cons44way) Track of the UCSC Tables were also obtained (Appendix I).

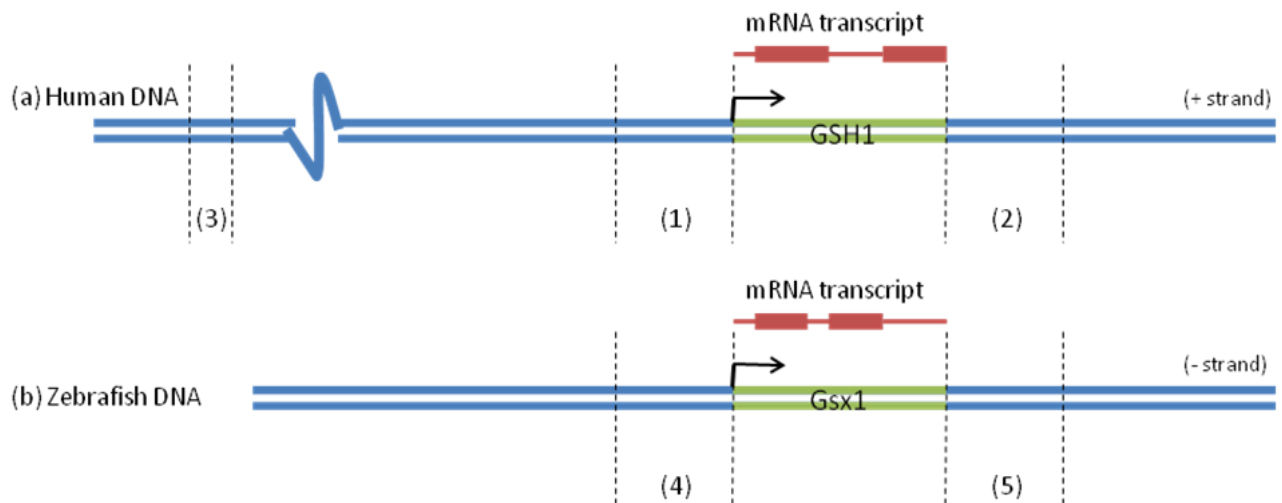


Figure 2.1.1. Regulatory regions analysed. Analysis of the non-coding regulatory region of *Gsx1* were broken down into 5 sub-regions, each regions being 1000 bp long (Regions 1, 2, 4 and 5), except for Region 3 which was only about 350 bp long (and about 1 Mbp upstream of *GSH1*). The thicker lines of the mRNA transcript are the exons.

## 2.2 BigFoot: Statistical alignment and phylogenetic footprinting software

BigFoot software package is an extension of a pre-existing algorithm that combines alignment and footprinting, using a MCMC approach to sample both sequence alignments and locations of slowly evolving regions. The BigFoot software package is downloadable from <http://www.stats.ox.ac.uk/~satija/BigFoot/>.

BigFoot models three evolutionary events, namely insertions, deletions and substitutions, to explain the evolutionary history between the sequences. Unlike traditional alignment algorithms which assumed that the rates of evolutionary events are identical along a sequence, the alignment model in BigFoot allowed rate heterogeneity by modeling the evolution of both quickly and slowly evolving regions. This alignment model is expressed as a pairwise HMM transducer – a conditionally normalized HMM representing the evolution of an ancestral (inputted) sequence into a descendent sequence. Further details on the model are presented in Appendix II.

To model the fast states and the slow states, BigFoot would scale down the branch lengths of the phylogenetic tree in the slow states, hence reducing the evolutionary time and the expected divergence of these regions. Different scaling factors for substitution events, and indel events, would be used, and these are the parameters to the model.

BigFoot employed a Bayesian MCMC sampling to estimate all the parameter distributions. User-input Gaussian or uniform priors can also be facilitated, to allow the user to tailor the analysis to their specific needs. Otherwise, uninformative, exponential priors with expectation 1 would be used, which allows the MCMC to freely estimate parameter distributions. The application of MCMC sampling is to converge the posterior distribution of the alignments, trees and evolutionary parameters to a prescribed distribution, because direct sampling of the posterior distribution would be impossible due to the high dimensional, complicated distribution form. Therefore, after convergence, samples from the Markov chain would be the correlated samples from the posterior distribution.

There are four components in the space where MCMC performed the random walk, (1) changing model parameters; (2) changing extended alignment, which refers to the distribution of conditional likelihoods at the internal nodes of the tree; (3) shifting the boundary of a n existing fast or slow region; and (4) creating a new (or deleting an old) fast or slow region.

Consequently, the Markov chain produced correlated samples from the posterior distribution of alignments, boundaries between fast and slow regions, and the evolutionary parameters. In BigFoot, the multiple sequence alignment is summarized by estimating the MPD (Maximum Posterior Decoding) alignment – an alignment that maximizes the product of posterior probabilities of alignment columns. These postprocessed samples would then appear in the Graphical User Interfaces (GUIs) implemented in the BigFoot package. When the analysis ended, all the footprinting posterior probabilities and the log-likelihood values were written into a text file.

### **2.3 Clustal X software**

Clustal X is a GUI for the ClustalW multiple sequence alignment program, which enables user to perform multiple sequence and profile alignments and analysis of results. The program is available from the Clustal Homepage ([www.clustal.org](http://www.clustal.org)) or European Bioinformatics Institute ftp server (<ftp://ftp.ebi.ac.uk/pub/software/>).

There were three main steps in performing a multiple sequence alignment with this program. Clustal X would first carry out a pairwise alignment, followed by the construction of a phylogenetic tree. In the last step, the phylogenetic tree is used to perform the multiple

alignment. All the three steps were done automatically by selecting “Do Complete Alignment”. Alternatively, a user-defined tree can also be used to create the phylogenetic tree in the second step.

After the multiple alignment process, Clustal X stored the phylogenetic tree in a file with an extension “.dnd” appended.

## **2.4 Analysis procedure**

Whole-genome alignment of a human or zebrafish region (i.e. Region 1 – 5) with other homologous species (according to the UCSC Conservation Track), in groups of 5 or 6 species, were entered into BigFoot in FASTA format, and then set to run. In addition, a phylogenetic tree in the Newick format, produced by ClustalX to describe the phylogenetic relationship between the inputted species, was also inserted.

Beforehand, phylogenetic trees of all the 44 species for each region were drawn using ClustalX, which resulted in 5 main trees being produced as a starting reference for the analyses (Appendix III). The group of species for each analysis were chosen based on their evolutionary distance from the main species of interest (i.e. human or zebrafish), since like other phylogenetic footprinting tools, BigFoot works best if the species being analysed are sufficiently evolutionarily distant, but not too distant.

The MCMC parameters used in BigFoot for each analysis were set to 1,000,000 Burn-In cycles and 500,000 MCMC sampling cycles, with a sampling rate of 5000. These values were deemed enough, since most of the time, the convergence of the MCMC sampling to the prescribed posterior distribution occurred within the 1,000,000 Burn-In cycles. However, if the log-likelihood trace of the test sampled still has not fluctuated steadily around a plateau, then the sample has not converged, and further test must be repeated using different MCMC parameters. In particular, a higher Burn-In value would be applied.

In the analysis of the dataset sequences, default prior settings of the alignment model parameters were the best option, because currently there is no known prior knowledge on the analyzed region available.

The conservation rate for each site which roughly corresponds to the probability that the column is slowly evolving (i.e. potential functional regulatory element) is written in a file

with an extension “.pred” appended. In the GUI plugin, these values were plotted as a red curve in the MPD alignment panel, where higher values indicate a greater posterior probability of purifying selection.

## 2.5 Comparative assessment

The dataset was also tried out in two other programs which have been commonly used to identify regulatory motifs in genomic sequences; phastCons and ConSite. Subsequently, results from BigFoot, phastCons and ConSite were compared and evaluated.

### *Comparison with phastCons*

PhastCons is one of the most widely used alignment-based phylogenetic footprinting tool, and indeed, this program is the basis of the conservation track in the UCSC genome browser. PhastCons is based on two-state phylogenetic HMM to identify conserved elements and produce conservation scores.

PhastCons analyses were made using web-interface version via UCSC genome browser instead of the command-line version. Therefore, the results produced were predicted from a single multiz alignment and does not incorporate indel information.

### *Comparison with ConSite*

ConSite is a web-based tool for finding *cis*-regulatory elements, where the predictions are based on the integration of binding site prediction generated with high-quality TF models chosen by the user and phylogenetic footprinting.

The profile models used for predicting the TFs in ConSite are drawn from the JASPAR database – an open-access, non-redundant collection of curated matrix-based TFBSs profiles for eukaryotes. The results were visualized either as a graphical conservation plot, as an annotated alignment, or in tabular form. The profiles are converted to log-scaled position weight matrices in order to evaluate possible binding sites in an input sequence, but since score ranges are unique for each model, the final score from ConSite is normalized.

### 3. RESULTS

#### 3.1 TFBSs prediction: The computational pipeline

The step-wise computational approach to TFBSs prediction in this study is summarized in Figure 3.1.1.

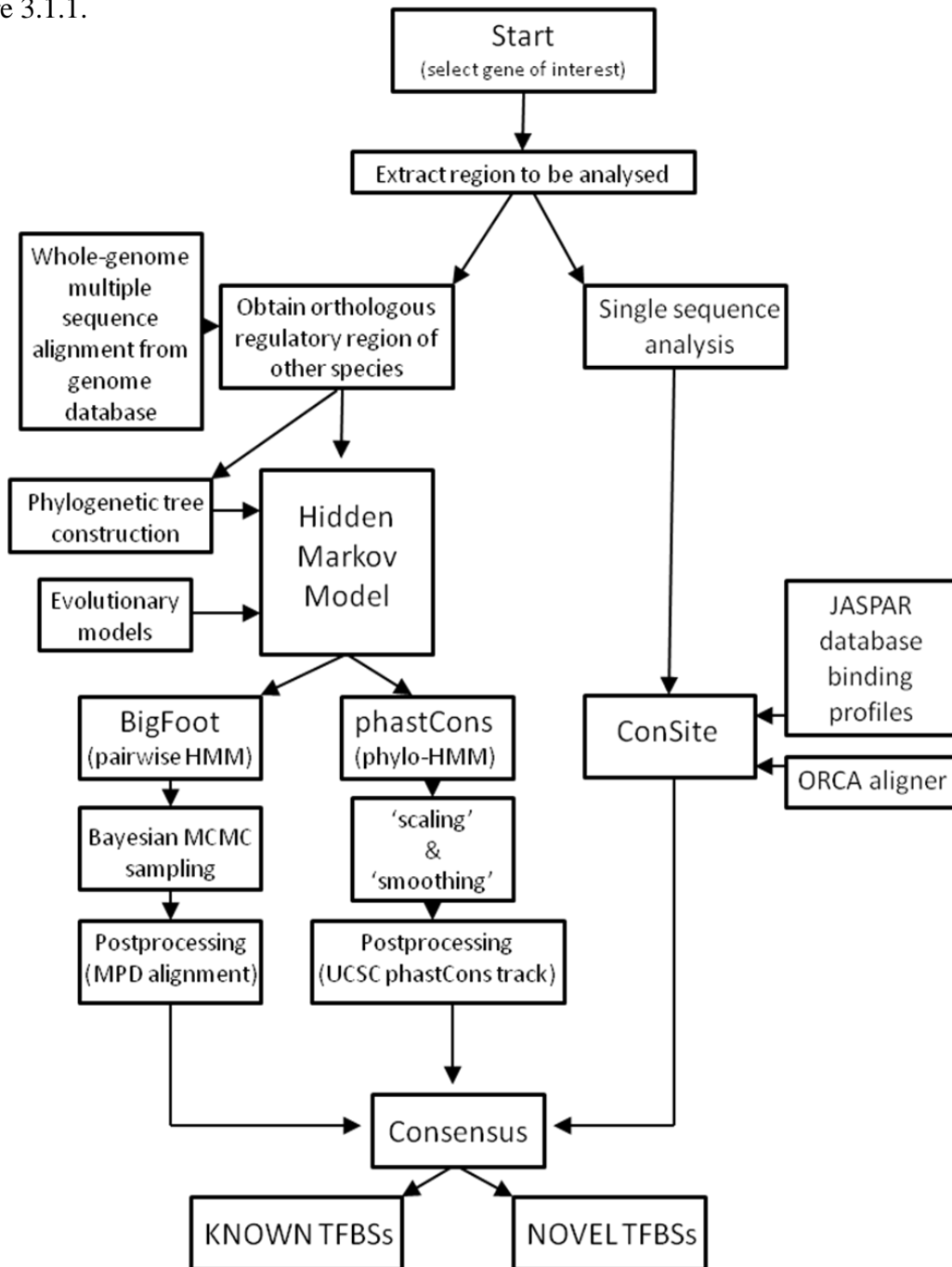


Figure 3.1.1 Strategy for computation based TFBSs prediction analysis. Common genome databases included UCSC Genome Browser, Ensembl Genome Browser and Entrez Genome database. In general, this study employed statistical alignment and phylogenetic footprinting to predict TFBSs. The ultimate test for the validity of predictions made by computational methods, however, is still *in vivo* experimental analysis.

### 3.2 BigFoot results

BigFoot had shown that getting the right combination of species for comparison is essential in detecting the slowly-evolving, conserved regulatory elements (example Figure 3.2.1). Additionally, preliminary tests using different MCMC parameters demonstrated that quite distinct outcome would occur if the sampling has not converged (example Figure 3.2.2).

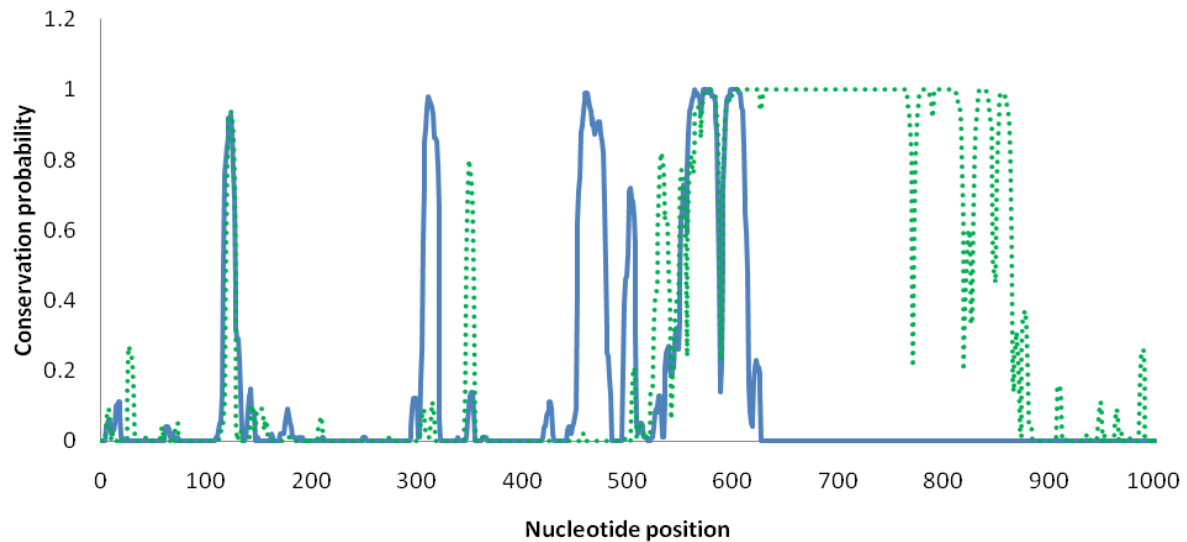


Figure 3.2.1 BigFoot output of two separate analyses of Region 1. Blue solid line is the result of comparing the human sequence with mouse, dog, cow and cat; while green dotted line is the result of comparing the human sequence with rhesus monkey, dog, cow and megabat (which has closer evolutionary distance according to Region 1 tree, Appendix 1).

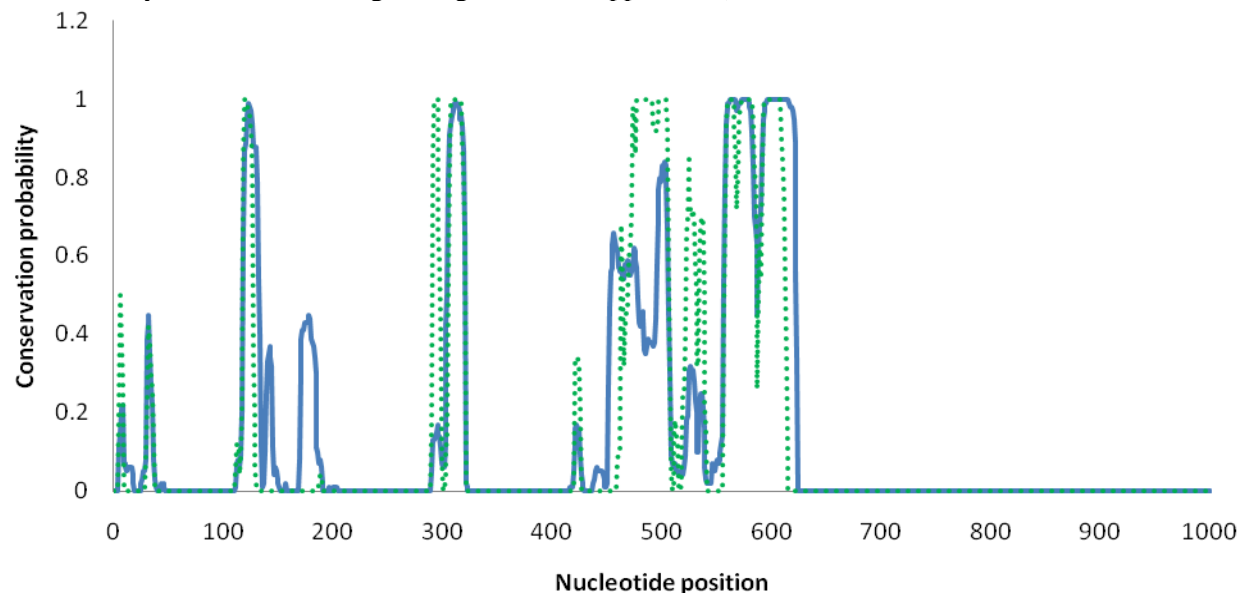


Figure 3.2.2 BigFoot output of two separate analyses of Region 1. In both analyses, the human sequence is compared against mouse, tenrec, cow and cat. Blue solid line is the result produced after MCMC sampling converged (Burn-In cycle of 1,000,000); green dotted line is an example of a result where the MCMC sampling has not converged (Burn-In cycle of 10,000).

The length of time it takes for BigFoot analysis varied - it can take between 3 hours to more than a day for each analysis on a standard desktop PC, depending on the length of sequence being analysed and the evolutionary distance between each species.

### **3.3 Promoter analysis of human and zebrafish Gsx1**

Regulatory element predictions of each region from BigFoot are compared to PhastCons and ConSite analyses of the same region (Figure 3.3.1; Figure 3.3.3 – Figure 3.3.5). TFBSs predicted with ConSite program used the setting for vertebrates, with minimum bit specificity of 7 bits and 95% TF score cutoff. However, only those with a score of greater than 9.0 are shown. The profiles of those forming consensus with BigFoot and phastCons results are also given.

In the BigFoot analysis, human Region 1 sequence was compared against homologous sequences from mouse, tenrec, cow and cat. This resulted in 6 distinctly conserved, slowly-evolving regions, found between - 400 to - 900 bp upstream of TSS. PhastCons detected 5 conserved regions, while ConSite predicted 9 known TFBSs. Out of the 9 TFBSs predicted by ConSite, 4 sites formed a consensus with the other two programs. The alignment produced by BigFoot for Region 1 is shown in Figure 3.3.2. Alignments from other regions are attached as Appendix IV.

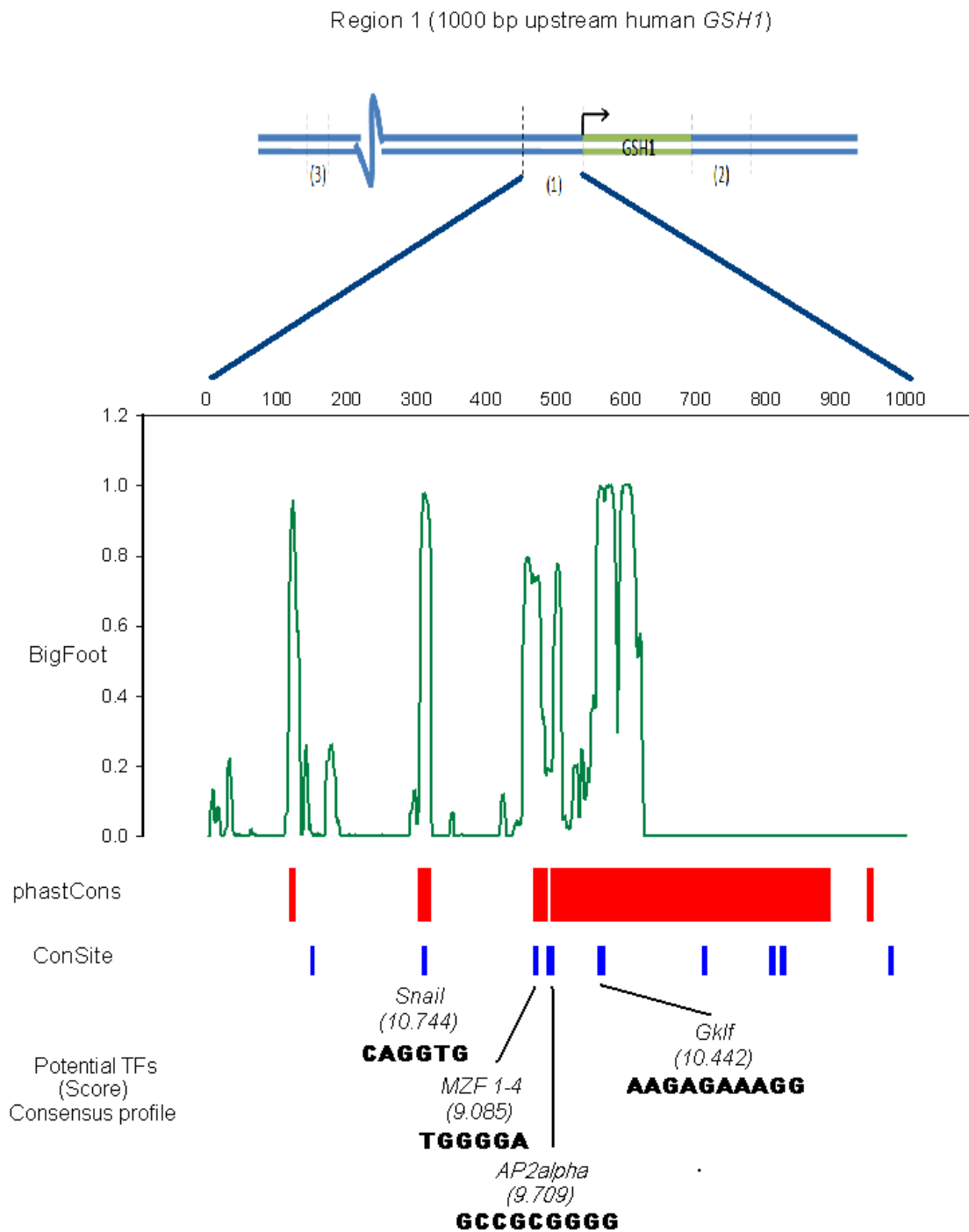


Figure 3.3.1 Regulatory elements prediction for Region 1. The BigFoot analysis produced 6 distinct peaks, found between -400 to -900 bp (upstream of TSS). PhastCons detected 5 conserved regions, while ConSite predicted 9 known TFBSs.



Human Region 2 sequence (Figure 3.3.3) was compared against homologous sequences from monkey, mouse lemur, dog, squirrel and elephant. The alignment produced by Bigfoot is fairly good, despite the numerous peaks shown which might be regarded as ‘noise’. PhastCons detected 6 conserved regions, but ConSite only managed to identify 4 known TFBSs from this region. Only 1 out of the 4 predicted sites formed a consensus with results from BigFoot and phastCons.

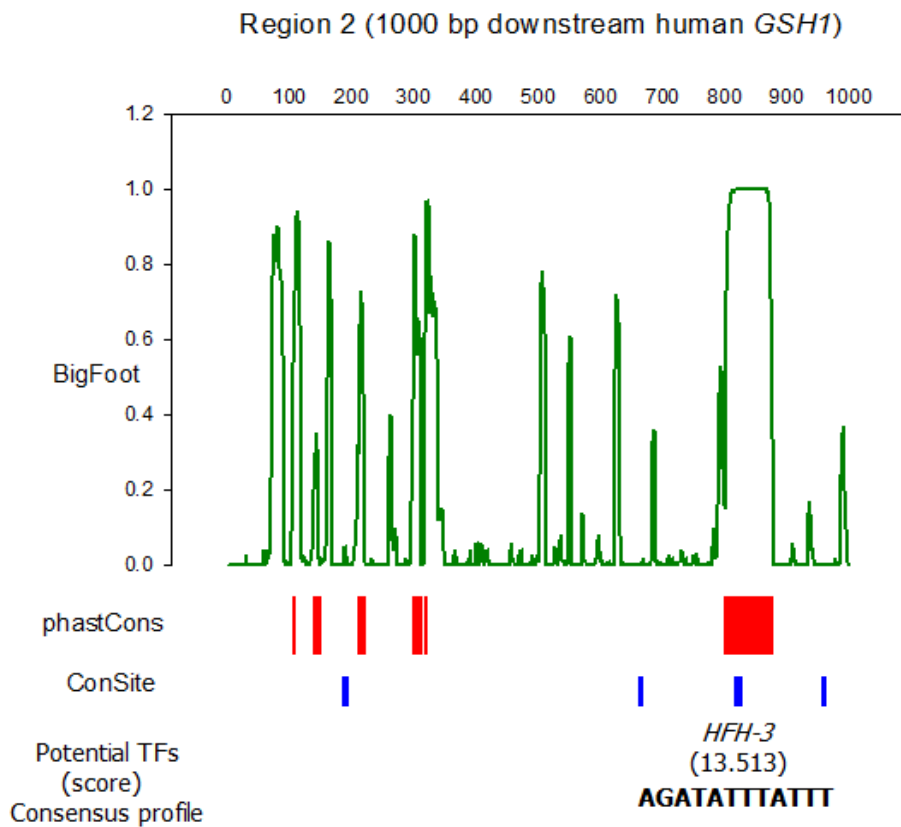


Figure 3.3.3 Regulatory elements prediction for Region 2. The BigFoot analysis detected 10 peaks with posterior probability of greater than 0.6; phastCons distinguished 6 conserved regions; while ConSite only managed to identify 4 known TFBSs from human Region 2.

Human Region 3 sequence (Figure 3.3.4) was compared against homologous sequences from cat, microbat, rabbit and horse, where BigFoot found 4 conserved, slow-evolving regions. Although this region has been annotated to interact with TF SP1 in a ChIP assay, there are no conserved regions detected by phastCons; while the 2 sites predicted by ConSite for this region are identified as c-ETS element with a score of 7.633 for both sites.

Region 3 ( 350 bp located 1 Mbp upstream human *GSH1*)

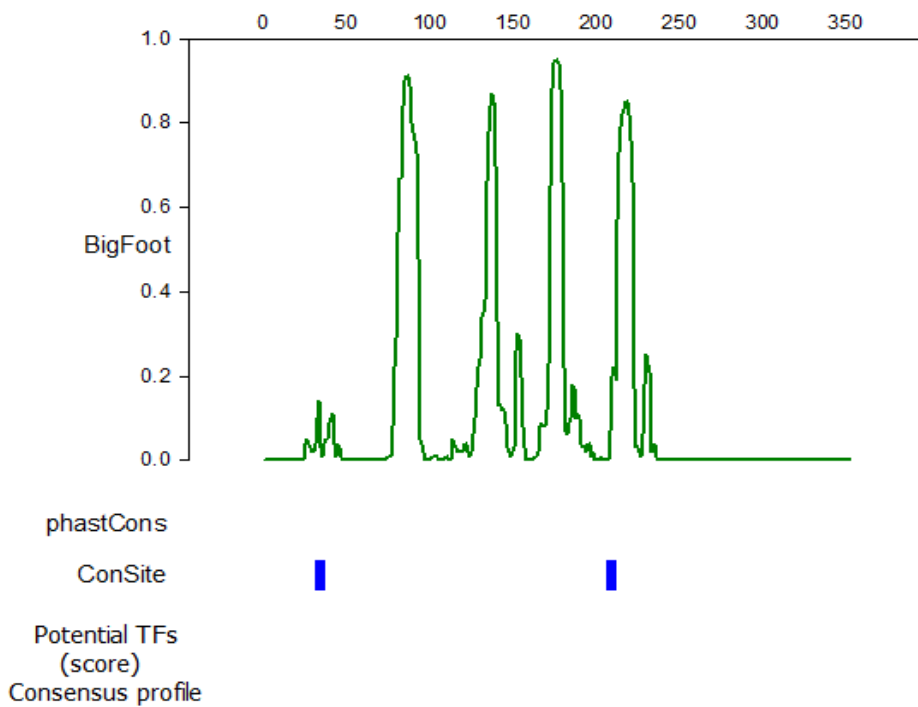


Figure 3.3.4 Regulatory elements prediction for Region 3. In human Region 3, the BigFoot analysis produced 4 distinct peaks. No conserved regions are detected by phastCons, and only 2 sites are predicted by ConSite.

Region 4 (1000 bp upstream zebrafish *Gsx1*)

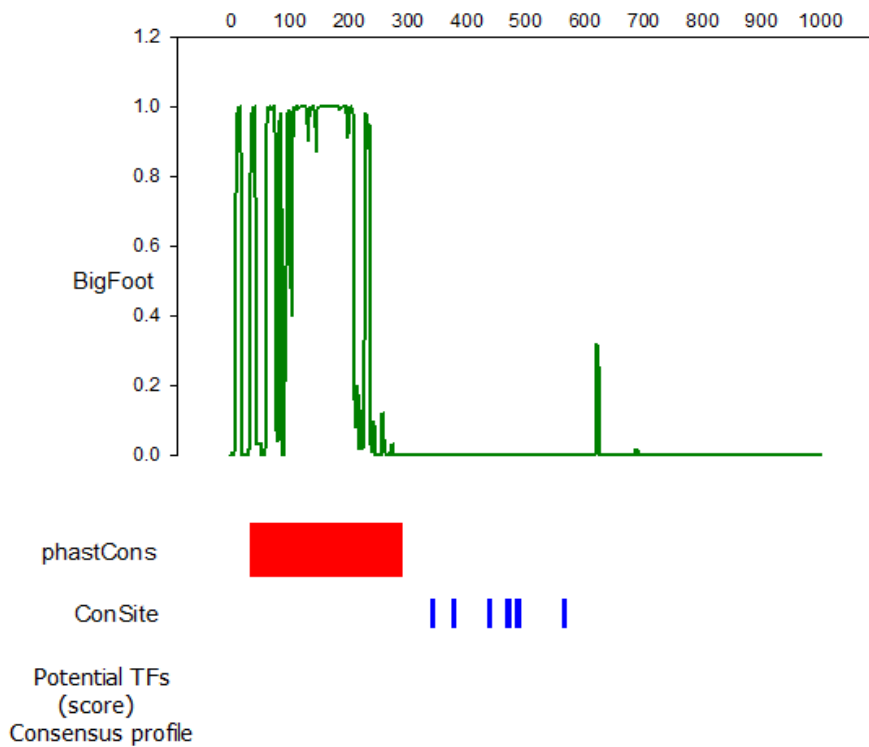


Figure 3.3.5 Regulatory elements prediction for Region 4. In the BigFoot analysis, all 5 conserved, slow-evolving regions are detected between - 1000 to - 700 bp. PhastCons detected 1 conserved region, while ConSite predicted 6 TFBSs.

Zebrafish Region 4 (Figure 3.3.5) sequence was compared against homologous sequences from fugu, tetraodon, human and mouse. This region is also detected by phastCons as one conserved chunk. However, the 6 sites identified by ConSite did not correspond to any of the regions suggested by BigFoot and phastCons.

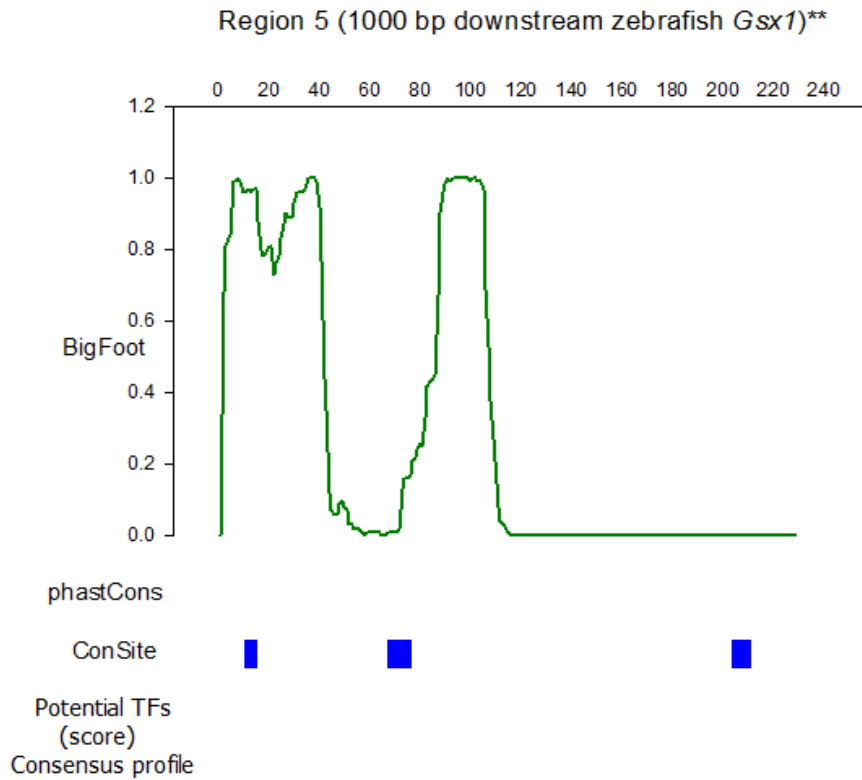


Figure 3.3.6 Regulatory elements prediction for Region 5. Analysis for zebrafish Region 5 sequence focused on the region + 445 to + 673 bp , downstream transcription stop site. BigFoot detected 2 main conserved regions which are not detected by phastCons. ConSite predicted 3 sites with known TFs for this ~ 230 bp-long region.

Zebrafish Region 5 sequence (Figure 3.3.6) was compared against homologous sequences from fugu, tetraodon, human and *Xenopus*. The 1000 bp-long region is cut down to ~230 bp since the homologous sequence from other species only matched the region + 445 to + 673 bp downstream of transcription stop site of zebrafish *Gsx1*. BigFoot detected 2 fairly big conserved regions which are not detected by phastCons at all, and one out of the three sites identified by ConSite is included in this conserved region.

### 3.4 Overrepresented TF binding motifs

There are 7 TFs predicted from ConSite which are found in both human and zebrafish sequences; *Ahr-ARNT*, *c-ETS*, *delta-EF1*, *Gklf*, *S8*, *SPI-1* and *TCF11-MafG*. ConSite TFs predictions for four other mammals (mouse, rhesus monkey, cat and cow) and two other fish (tetraodon and fugu) were also made for comparison (Table 3.4.1). For the mammals' category, sequences homologous to human Region 1 (i.e. from UCSC cons44way track) were used; while sequences homologous to zebrafish Region 4 were employed in the analyses of the fish category. The outcome is summarized in Table 3.4.2.

Table 3.4.1 TFBSs predicted by ConSite program. Minimum bit specificity of 7 bits and 95% TF score cutoff was used in the single-sequence analysis of the promoter region of the 8 species.

Predicted TFs	Human	Mouse	Rhesus monkey	Cow	Cat	Zebrafish	Fugu	Tetraodon
Ahr-ARNT	✓	✓	✓		✓	✓		
AP2alpha	✓	✓	✓	✓				
c-ETS	✓	✓	✓	✓	✓	✓		
deltaEF1	✓	✓	✓	✓	✓	✓		
Elk-1			✓					
Gklf	✓	✓			✓	✓		
MZF_1-4	✓	✓	✓	✓	✓		✓	✓
MZF_5-13					✓	✓		
Nkx				✓				
Pbx						✓		
S8	✓		✓	✓		✓		
Sox-5						✓		
SPI-1	✓	✓	✓	✓	✓	✓	✓	✓
SPI-B	✓	✓	✓	✓				
SRY						✓		
TCF11-MafG	✓		✓			✓		
Yin-Yang	✓				✓			

Table 3.4.2 Summary of overrepresented TF binding motifs as predicted by ConSite program.

Category	Species	No. of TFBSs predicted by ConSite <sup>(a)</sup>	Overrepresented TF profiles within each category <sup>(b)</sup>	Overrepresented TF profiles between each category <sup>(b)</sup>
Mammals	Human	24	<i>Ahr-ARNT</i> <i>AP2-alpha</i> <b><i>c-ETS</i></b> <b><i>delta-EF1</i></b> <i>Gklf</i> <b><i>MZF 1-4</i></b> <i>S8</i> <b><i>SPI-1</i></b> <i>SPI-B</i>	<i>Ahr-ARNT</i> <i>AP2-alpha</i> <i>c-ETS</i> <i>delta-EF1</i> <i>Gklf</i> <i>MZF 1-4</i> <i>S8</i> <b><i>SPI-1</i></b> <i>SPI-B</i>
	Mouse	22		
	Rhesus monkey	24		
	Cat	15		
	Cow	13		
Fish	Zebrafish	23	<b><i>SPI-1</i></b> <i>MZF 1-4</i>	<b><i>SPI-1</i></b> <i>SPI-B</i>
	Tetraodon <sup>(c)</sup>	2		
	Fugu <sup>(c)</sup>	2		

- (a) Some TFBSs can be identified by more than one TF, but these sites are regarded as 1 TFBS in the count.
- (b) Only TF elements which are detected in more than half of the species in that category are listed. TFs which are identified in all the species for that category are in **bold**.
- (c) The sequences from tetraodon and fugu (homologous to zebrafish Region 4) obtained are only 268 bp long.

### 3.5 Potential novel motifs

There are 8 sub-regions within the 5 main regions where a consensus is formed between BigFoot and phastCons (Table 3.5.1). These sub-regions have no known TF predicted by ConSite, but we should not exclude the possibility that these sub-regions might harbour potential novel motifs. However, only elements greater than 6 bp and less than 25 bp are attained.

Table 3.5.1 Potential novel binding sites for the 5 regions. Regions which formed a consensus between BigFoot and phastCons, and with length of 6 – 25 bp, are listed as potential novel binding motifs.

Location	Potential binding motif
<b>Region 1</b>	
120 - 128	aaCAAAGac
<b>Region 2</b>	
211 - 221	GCCAACTcaa
300 - 312	GCGCCTTGCTGGC
<b>Region 4</b>	
9 - 17	TGGTGCTGA
34 - 42	CTTAAATAG
59 - 76	CCCATGTGACGGTTACGC
95 - 101	CTCAATA
226 - 237	CTGCCCTTCTC

## 4. DISCUSSION

### 4.1 The phylogenetic footprinting approach

Phylogenetic footprinting has become widely exploited approach in refining searches for TFBSs because of the significantly better predictions produced by the phylogenetic filtering procedure. Indeed, phylogenetic footprinting managed to dramatically reduce false-positive predictions of TFBSs as mentioned earlier and has been used in several studies, in particular, as a primary research tool in identifying regulatory elements (Lenhard *et al.*, 2003; Siepel *et al.*, 2005).

However, this comparative approach is often complicated by several factors. First, despite ample evidence that conserved regions do often contain functional regulatory motifs (as reviewed in Maston *et al.*, 2006), there are examples which suggest otherwise. For instance, (Balhoff and Wray, 2005) found that TFBSs in the *endo16* gene promoter are no more conserved than the surrounding, non-functional sequence when comparative sequence analysis is made. Indeed, we are still unable to precisely correlate conserved and functional regions, due to the presence of a large amount of highly conserved non-coding sequences in the human genome with unknown function (Maston *et al.*, 2006). Secondly, not all TFBSs are conserved among species. As mentioned before, TFs bind to degenerate sequence motifs, so perfect sequence conservation of an element is not required. This means that the same TF might bind to variant TFBSs that are present in different species; and although gene expression patterns might still be conserved across species, redundancy of regulatory elements can allow a single element to be gained or lost without affecting the overall gene expression of the gene. In other words, the use of phylogenetic footprinting to detect functional regulatory elements will be a waste of time if the regulatory element in question is not conserved across different species (Maston *et al.*, 2006).

Additionally, Prakash and Tompa (2005) identified two other issues when applying phylogenetic footprinting particularly to vertebrates. First, the idiosyncrasies of vertebrate genome that make the process of obtaining a reliable set of orthologous promoter region more difficult than with simpler genomes. It is difficult to identify orthologous genes among vertebrates as illustrated by Prakash and Tompa (2005); 16% of the genes involved in their study showed inconsistency in Ensembl's homology mappings (Prakash and Tompa, 2005). When a set of orthologous genes has been determined for every species, there is also the issue

of extracting the right regions for analysis. Most groups tend to extract and analyze the regions just upstream of the annotated translation start sites (e.g.(Elemento and Tavazoie, 2005). However, in the case of vertebrate genome, it is common that the annotated translation start sites of orthologous genes are not orthologous positions, due to loss of the first exon in some species, errors in annotation or lack of experimental evidence for start sites (Prakash and Tompa, 2005). In addition, the genomic distance between transcription and translation start sites can be very large, at such a point 1000 bp upstream of the translation start site might lie downstream of the TSS.

Hence, to avoid the issue of orthologs, sequence of homologous regions (i.e. homologous to the main species of interest, human or zebrafish) from other species were obtained via whole-genome multiple alignment of 44 vertebrates species using to the Conservation (cons44way) track in UCSC Genome Browser. In obtaining the sequences, initially, pairwise alignments with the human genome were generated for each species. Then an additional filtering step was introduced in the generation of the 44-way conservation track to reduce the number of paralogs, pseudogenes and suspected alignments. The resulting best-in-genome pairwise alignments were progressively aligned using multiz program (Blanchette *et al.*, 2004), according to a pre-determined tree topology, to produce multiple alignments (UCSC Conservation Track website description).

This approach is deemed acceptable in this study because complete annotation of the orthologous regulatory region of other species is not necessary for BigFoot to operate. Surely some bona fide TFBSs that are not well conserved between species might be present, and this approach will miss them. However, this issue lies within the basis of phylogenetic footprinting that cannot be avoided. Nevertheless, it is clear that the issue of homologs is important in detecting sites when using the comparative approach and that substantial amount of work on complete and accurate annotation of common sequence databases is required.

## **4.2 Statistical implementation of comparative approach**

Many of the early successes of the vertebrate comparative genomics were sequence comparisons between human and mouse (reviewed in Stone *et al.*, 2005), where their evolutionary distance is sufficient but not excessive. The ability of comparative approach to detect conserved functional elements, which are under purifying selection, is intimately tied to the neutral rate of evolution (Stone *et al.*, 2005). Although orthologous regulatory elements

might have evolved slowly due to selective constraints and can be detected by their sequence similarity, ancestral sequences that have evolved slowly by chance are indistinguishable from their functional counterparts. Additionally, vast majority of functional elements, although under purifying selection, can stochastically accumulate substitutions at a vital fraction of the neutral rate (Stone *et al.*, 2005). By chance, some elements might have changed so much that they fall short of the conservation threshold required for detection. Therefore, specificity, which quantifies the extent to which neutral sequence is misidentified as functional, and sensitivity, which quantifies the extent to which functional sequence is misidentified as neutral because of accumulated evolutionary change, are important parameters that need to be taken into account when defining new models to effectively discriminate neutrally evolving sequence from its functional counterpart (Stone *et al.*, 2005).

Various models have been proposed to counter such problems, but one that has shown its credibility and has been adapted to many sequence evolution related problems is the hidden Markov model, first proposed by (Felsenstein and Churchill, 1996) to model substitution rate correlation along the genome. However, it was (Qian and Goldstein, 2003) who first introduced a method to incorporate phylogenetic information directly into HMMs and demonstrated that the resulting model performs better than most of the multiple sequence-based methods at that time. Indeed, when Siepel and Haussler (2006) investigated further the importance of phylogeny in HMMs, they found that the performance of phylogenetic HMM rapidly improves the nucleotide-level sensitivity and specificity as the number of species ( $n$ ) increases; about 98% sensitivity and specificity achieved by  $n = 2$ , and 99% sensitivity and specificity achieved by  $n = 5$ . Without the phylogenetic information, on the other hand, decline in both sensitivity and specificity was observed (Siepel and Haussler, 2005). Therefore, mathematical development of HMM could hopefully lead to better phylogenetic footprinting algorithms, and eventually better TFBSs detection programs.

### **4.3 BigFoot: strength and limitation**

Most likely, better models and programs will be designed as total number of species with available genomic sequence increase. However, for this study, BigFoot, in particular, has exhibited superior performance compared to other existing tools.

Running BigFoot has shown that it might not be as user-friendly as other programs such as VISTA (Frazer *et al.*, 2004), especially to the likes of biologists who used bioinformatic tools

as pre-research tests, since each analysis is time-consuming and the output quality depends highly on the quality of the inputted sequences. BigFoot analysis results can be quite challenging to interpret, hence running several preliminary tests are recommended, especially when no prior knowledge on the region of interest is available.

However, BigFoot has shown that it is an advanced program with realistic models that are able to accommodate and account for the spatial rate variation and correlation in genome evolution. In other words, BigFoot managed to reflect the nature of real data more accurately. Indeed, in most of the analyses, BigFoot has shown greater sensitivity and specificity compared to phastCons, although expression data are further required to verify the results. Moreover, there is still room for improvement for BigFoot in terms of user-interface and it is probable that the future versions of this program will only get better.

#### **4.4 Regulatory elements of vertebrate *Gsx1***

There are very few studies on vertebrate *Gsx1* regulation, and where they do exist, most concentrated on the ParaHox cluster as a whole. No studies have examined the regulation of vertebrate *Gsx1* particularly, and the closest comparison that can be made is the study of retinoic acid (RA) regulation of ParaHox genes in the invertebrate chordate amphioxus (*B. floridae*) by Osborne *et al.* (2009). Osborne *et al.* (2009) found in their study that expression of all three ParaHox genes (i.e. *AmphiGsx*, *AmphiXlox* and *AmphiCdx*) is clearly modified by RA treatment, and hence hypothesised a direct regulation of ParaHox genes by RA. They identified a RA response element (RARE) consensus sequence upstream of *AmphiGsx* which they labelled as DR5e. However, when using a combination of electrophoretic mobility shift assay (EMSA) and heterologous cell culture transactivation experiments for *in vitro* test, no DR5e element was detected (Osborne *et al.*, 2009).

Therefore, this study is the first to address the promoter analysis of vertebrate *Gsx1*. The computational approach had found several potential regulatory elements of vertebrate *Gsx1* and suggested a list of transcription factor candidates. One of particular interest is the MZF 1-4 element. MZF protein is a C2H2 zinc finger protein first identified in developing myeloid cells (Hromas *et al.*, 1991). In human *GSH1*, this element was detected in Region 1 with relatively high probability/score by all the three programs, and further ConSite analysis has shown that MZF 1-4 element can be identified in the defined regulatory region of all species except zebrafish (Table 3.4.1). MZF 1-4 falls under general TF class as it is known to be

involved in expression and regulation of several transcription initiation complex genes (e.g. (Kim *et al.*, 2001; Yuan *et al.*, 2006) , so it is likely that *Gsx1* is one of the “realisator genes”.

Another interesting element predicted via the computational approach is the Gklf element, also detected by the three programs in human *GSH1* Region 1. Gklf is also detected in zebrafish by ConSite, but it does not correspond to any conserved sites by phastCons. Gklf is also known as Kruppel-like factor 4 (gut) or KLF4, and can act either as transcriptional activator or repressor depending on the promoter context. Gklf has been shown to interact with CREB binding protein (Geiman *et al.*, 2000), which in turn, plays a critical role in embryonic development.

Three of the elements also identified in the promoter region of vertebrate *Gsx1* are known to interact with the ETS domain, a winged helix-turn-helix structure. TFs c-ETS, SPI-1 and SPI-B belong to the ETS TFs family, which is one of the largest families of TFs unique to metazoans. c-ETS is known to cooperate with c-Fos and c-Jun, two well-known early and intermediate response TFs (Wasylyk *et al.*, 1990), so it might be that the early presence of c-ETS is involved in the regulation of *Gsx1* gene.

However, precaution must be taken when deciphering results from computational analysis. For instance, analysis of predicted TFs from ConSite suggested that SPI-1 is present on the regulatory region of all the species involved. However, SPI-1 are known, in humans at least, to be specifically involved in the differentiation or activation of macrophages or B-cells (Galson *et al.*, 1993).

Of course, these predicted TFs are suggested because they correspond to the known TFBSs profiles in JASPAR database. The computational approach also predicts novel binding motifs which can potentially be involved in the regulation of *Gsx1*. However, whether these motifs are *cis*-elements that regulate *Gsx1* gene *in vitro* and *in vivo* in vertebrates remain to be established in experiments.

## **4.5 Future work**

In this study, BigFoot results were compared with results from other programs using almost similar approach because no experimentally verified data was available. If there had been experimental inputs available, we could have better evaluation on the reliability of BigFoot in detecting conserved regulatory elements. On the other hand, the computational promoter

analysis of human *GSH1* and zebrafish *Gsx1* now can form an additional layer of a priori knowledge which can influence the design of laboratory investigations.

Additionally, the poor quality of genomic data available on vertebrate *Gsx1*, particularly for zebrafish, constrained the full capacity of computational approach. For instance, only four other species were found to contain homologous region with Region 5 (1000 bp downstream zebrafish *Gsx1* transcription stop site), where two of them are poorly aligned. BigFoot, on the other hand, will produce poor nucleotide resolution if the number of species used for comparison is less than four. If there had been more time for this study, we could have mined for better data from genomic sequence databases other than UCSC Genome Browser, and probably apply the techniques used by (Prakash and Tompa, 2005) to obtain orthologous region to Region 5 in other species as comparison.

Future work might also involve using additional information, such as results from other promoter and TFBSs prediction programs; and structural data of vertebrate *Gsx1*, particularly in the context of chromatin structure, which is known to play a critical role in gene regulation. When full information is not available, these data might help in formulating better criteria for annotation.

## 5. CONCLUSION

Rapid development in statistical analysis and model building, together with the accumulation of genomic data and expression data make it increasingly essential to remind ourselves that *in vivo* validation needs to be combined with the computational approach in order to understand transcriptional regulatory processes. In the field of phylogenetic footprinting, the bioinformatic tools are sophisticated enough for analysis with significant results to be produced.

Application of the computational approach to analyse regulatory regions of human and zebrafish *Gsx1* enable systematic predictions to be made, which can be used as a priori when designing the next stage of research – the experimental validation. With a consensus between three programs (BigFoot, phastCons and ConSite), this study has predicted five TFBSs in human *GSH1* regulatory region, but none in zebrafish *Gsx1* regulatory region.

Of course, this study alone would not suffice in offering further insights as to whether the ParaHox cluster are maintained in chordates as a result of selective constraints – genomic mapping in amphioxus indicated that the ParaHox cluster covers a region of ~ 30 kb (Brooke *et al.*, 1998). However, it has shown that we can now harness the full advantage of computational analysis in promoter prediction and TFBSs identification.

## REFERENCES

- Balhoff, J. P. and Wray, G. A. (2005). Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. *Proceedings of the National Academy of Sciences of the United States of America* 102 (24), pp.8591-8596.
- Birney, E. *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, pp.799-816.
- Bishop, M. J. and Thompson, E. A. (1986). MAXIMUM-LIKELIHOOD ALIGNMENT OF DNA-SEQUENCES. *Journal of Molecular Biology* 190 (2), pp.159-165.
- Blanchette, M. *et al.* (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* 14 (4), pp.708-715.
- Blanchette, M., Schwikowski, B. and Tompa, M. (2002). Algorithms for phylogenetic footprinting. *Journal of Computational Biology* 9 (2), pp.211-223.
- Brooke, N. M., Garcia-Fernandez, J. and Holland, P. W. H. (1998). The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 392 (6679), pp.920-922.
- Carroll, S. B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. *Cell* 101 (6), pp.577-580.
- Cartharius, K. *et al.* (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 21 (13), pp.2933-2942.
- Cohen, R. B., Sheffery, M. and Kim, C. G. (1986). PARTIAL-PURIFICATION OF A NUCLEAR-PROTEIN THAT BINDS TO THE CCAAT BOX OF THE MOUSE ALPHA-GLOBIN GENE. *Molecular and Cellular Biology* 6 (3), pp.821-832.
- Cooper, G. M. and Brown, C. D. (2008). Qualifying the relationship between sequence conservation and molecular function. *Genome Research* 18 (2), pp.201-205.
- Dimas, A.S. *et al.* (2009). Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science* DOI: 10.1126/science.1174148
- Elemento, O. and Tavazoie, S. (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biology* 6 (2).
- Felsenstein, J. and Churchill, G. A. (1996). A hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* 13 (1), pp.93-104.
- Ferrier, D. E. *et al.* (2005). The chordate ParaHox, cluster. *Current Biology* 15 (20), pp.R820-R822.
- Ferrier, D. E. K. and Minguillon, C. (2003). Evolution of the Hox/ParaHox gene clusters. *International Journal of Developmental Biology* 47 (7-8), pp.605-611.
- Frazer, K. A. *et al.* (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Research* 32, pp.W273-W279.
- Galson, D. L. *et al.* (1993). MOUSE BETA-GLOBIN DNA-BINDING PROTEIN B1 IS IDENTICAL TO A PROTOONCOGENE, THE TRANSCRIPTION FACTOR SPI-1/PU.1, AND IS RESTRICTED IN EXPRESSION TO HEMATOPOIETIC-CELLS AND THE TESTIS. *Molecular and Cellular Biology* 13 (5), pp.2929-2941.
- Geiman, D. E. *et al.* (2000). Transactivation and growth suppression by the gut-enriched Kruppel-like factor (Kruppel-like factor 4) are dependent on acidic amino acid residues and protein-protein interaction. *Nucleic Acids Research* 28 (5), pp.1106-1113.
- Greally, J. M. (2007). Genomics - Encyclopaedia of humble DNA. *Nature* 447 (7146), pp.782-783.

- Hein, J. *et al.* (2000). Statistical alignment: Computational properties, homology testing and goodness-of-fit. *Journal of Molecular Biology* 302 (1), pp.265-279.
- Hromas, R. *et al.* (1991). A RETINOIC ACID-RESPONSIVE HUMAN ZINC FINGER GENE, MZF-1, PREFERENTIALLY EXPRESSED IN MYELOID CELLS. *Journal of Biological Chemistry* 266 (22), pp.14183-14187.
- Illes, J. C., Winterbottom, E. and Isaacs, H. V. (2009). Cloning and Expression Analysis of the Anterior ParaHox Genes, Gsh1 and Gsh2 From *Xenopus tropicalis*. *Developmental Dynamics* 238 (1), pp.194-203.
- Jukes, T. H. and Cantor, C. R. (1969). EVOLUTION OF PROTEIN MOLECULES. *Munro, H. N. (Edited by). Mammalian Protein Metabolism. Vol. Iii. Xvii + 571p. Illus. Academic Press: New York, N.Y., U.S.A.*, pp.21-132.
- Kim, K. W. *et al.* (2001). Genomic structure of human GM3 synthase gene (hST3Gal V) and identification of mRNA isoforms in the 5'-untranslated region. *Gene* 273 (2), pp.163-171.
- Lenhard, B. *et al.* (2003). Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2 (2), p.13.
- Li, H. *et al.* (1996). Gsh-1, an orphan Hox gene, is required for normal pituitary development. *Embo Journal* 15 (4), pp.714-724.
- Lin, J. J.-C. *et al.* (2007). Characterization of cis-regulatory elements and transcription factor binding: gel mobility shift assay. *Methods Mol Biol* 366, pp.183-201.
- Maston, G. A., Evans, S. K. and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics* 7, pp.29-59.
- Mattick, J. S. (2003). The human genome and the future of medicine. *Medical Journal of Australia* 179 (4), pp.212-216.
- Mulley, J. F., Chiu, C. H. and Holland, P. W. H. (2006). Breakup of a homeobox cluster after genome duplication in teleosts. *Proceedings of the National Academy of Sciences of the United States of America* 103 (27), pp.10369-10372.
- Osborne, P. W. *et al.* (2009). Differential regulation of ParaHox genes by retinoic acid in the invertebrate chordate amphioxus (*Branchiostoma floridae*). *Developmental Biology* 327 (1), pp.252-262.
- Prabhakar, S. *et al.* (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Research* 16 (7), pp.855-863.
- Prakash, A. and Tompa, M. (2005). Discovery of regulatory elements in vertebrates through comparative genomics. *Nature Biotechnology* 23 (10), pp.1249-1256.
- Qian, B. and Goldstein, R. A. (2003). Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins-Structure Function and Genetics* 52 (3), pp.446-453.
- Qiu, P. (2003). Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. *Biochemical and Biophysical Research Communications* 309 (3), pp.495-501.
- Roth, F. P. *et al.* (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology* 16 (10), pp.939-945.
- Ruvkun, G. and Hobert, O. (1998). The taxonomy of developmental control in *Caenorhabditis elegans*. *Science* 282 (5396), pp.2033-2041.
- Sandelin, A., Wasserman, W. W. and Lenhard, B. (2004). ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Research* 32, pp.W249-W252.
- Satija, R., Pachter, L. and Hein, J. (2008). Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics* 24 (10), pp.1236-1242.

- Siepel, A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15 (8), pp.1034-1050.
- Stone, E. A., Cooper, G. M. and Sidow, A. (2005). Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annual Review of Genomics and Human Genetics* 6, pp.143-164.
- Takai, D. and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America* 99 (6), pp.3740-3745.
- Thorne, J. L., Kishino, H. and Felsenstein, J. (1991). AN EVOLUTIONARY MODEL FOR MAXIMUM-LIKELIHOOD ALIGNMENT OF DNA-SEQUENCES. *Journal of Molecular Evolution* 33 (2), pp.114-124.
- Venter, J. C. *et al.* (2001). The sequence of the human genome. *Science* 291 (5507), pp.1304-+.
- Wang, B. *et al.* (2009). Ascl1 is a required downstream effector of Gsx gene function in the embryonic mouse telencephalon. *Neural Development* 4.
- Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* 5 (4), pp.276-287.
- Waslyk, B. *et al.* (1990). THE C-ETS PROTOONCOGENES ENCODE TRANSCRIPTION FACTORS THAT COOPERATE WITH C-FOS AND C-JUN FOR TRANSCRIPTIONAL ACTIVATION. *Nature* 346 (6280), pp.191-193.
- Yuan, Z. F. *et al.* (2006). Genomic organization, promoter activity, and expression of the human choline transporter-like protein 1. *Physiological Genomics* 26 (1), pp.76-90.
- Zhang, M. Q. (1998). Identification of human gene core promoters in silico. *Genome Research* 8 (3), pp.319-326.
- Zhang, Z. and Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol* 2 (2), p.11.

Wheeler

Haussman?

Dimas, A.S. *et al.* (2009). Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. *Science* DOI: 10.1126/science.1174148

# APPENDICES

## APPENDIX I

### Species involved

*Genome assemblies included in the 44-way Conservation track.*

Organism	Species	Release date	UCSC version	Alignment type
Human	<i>Homo sapiens</i>	Mar 2006	hg18	reference species
Alpaca	<i>Vicugna pacos</i>	Jul. 2008	vicPac1	Reciprocal Best
Armadillo	<i>Dasypus novemcinctus</i>	Jul. 2008	dasNov2	Reciprocal Best
Bushbaby	<i>Otolemur garnettii</i>	Dec. 2006	otoGar1	Reciprocal Best
Cat	<i>Felis catus</i>	Mar. 2006	felCat3	Reciprocal Best
Chicken	<i>Gallus gallus</i>	May 2006	galGal3	Syntenic Net
Chimp	<i>Pan troglodytes</i>	Mar. 2006	panTro2	Syntenic Net
Cow	<i>Bos taurus</i>	Oct. 2007	bosTau4	Syntenic Net
Dog	<i>Canis lupus familiaris</i>	May 2005	canFam2	Syntenic Net
Dolphin	<i>Tursiops truncatus</i>	Feb. 2008	turTru1	Reciprocal Best
Elephant	<i>Loxodonta africana</i>	Jul. 2008	loxAfr2	Reciprocal Best
Fugu	<i>Takifugu rubripes</i>	Oct. 2004	fr2	MAF Net
Gorilla	<i>Gorilla gorilla gorilla</i>	Oct. 2008	gorGor1	Reciprocal Best
Guinea Pig	<i>Cavia porcellus</i>	Feb. 2008	cavPor3	Syntenic Net
Hedgehog	<i>Erinaceus europaeus</i>	June 2006	eriEur1	Reciprocal Best
Horse	<i>Equus caballus</i>	Sep. 2007	equCab2	Syntenic Net
Kangaroo rat	<i>Dipodomys ordii</i>	Jul. 2008	dipOrd1	Reciprocal Best
Lamprey	<i>Petromyzon marinus</i>	Mar. 2007	petMar1	MAF Net
Lizard	<i>Anolis carolinensis</i>	Feb. 2007	anoCar1	Reciprocal Best
Marmoset	<i>Callithrix jacchus</i>	June 2007	calJac1	Reciprocal Best
Medaka	<i>Oryzias latipes</i>	Oct. 2005	oryLat2	MAF Net
Megabat	<i>Pteropus vampyrus</i>	Jul. 2008	pteVam1	Reciprocal Best

Microbat	<i>Myotis lucifugus</i>	Mar. 2006	myoLuc1	Reciprocal Best
Mouse	<i>Mus musculus</i>	July 2007	mm9	Syntenic Net
Mouse lemur	<i>Microcebus murinus</i>	Jun. 2003	micMur1	Reciprocal Best
Opossum	<i>Monodelphis domestica</i>	Jan. 2006	monDom4	Syntenic Net
Orangutan	<i>Pongo pygmaeus abelii</i>	July 2007	ponAbe2	Syntenic Net
Pika	<i>Ochotona princeps</i>	Jul. 2008	ochPri2	Reciprocal Best
Platypus	<i>Ornithorhynchus anatinus</i>	Mar. 2007	ornAna1	Reciprocal Best
Rabbit	<i>Oryctolagus cuniculus</i>	May 2005	oryCun1	Reciprocal Best
Rat	<i>Rattus norvegicus</i>	Nov. 2004	rn4	Syntenic Net
Rhesus	<i>Macaca mulatta</i>	Jan. 2006	rheMac2	Syntenic Net
Rock hyrax	<i>Procavia capensis</i>	Jul. 2008	proCap1	Reciprocal Best
Shrew	<i>Sorex araneus</i>	June 2006	sorAra1	Reciprocal Best
Sloth	<i>Choloepus hoffmanni</i>	Jul. 2008	choHof1	Reciprocal Best
Squirrel	<i>Spermophilus tridecemlineatus</i>	Feb. 2008	speTri1	Reciprocal Best
Stickleback	<i>Gasterosteus aculeatus</i>	Feb. 2006	gasAcu1	MAF Net
Tarsier	<i>Tarsier syrichta</i>	Aug. 2008	tarSyr1	Reciprocal Best
Tenrec	<i>Echinops telfairi</i>	July 2005	echTel1	Reciprocal Best
Tetraodon	<i>Tetraodon nigroviridis</i>	Feb. 2004	tetNig1	MAF Net
TreeShrew	<i>Tupaia belangeri</i>	Dec. 2006	tupBel1	Reciprocal Best
X. tropicalis	<i>Xenopus tropicalis</i>	Aug. 2005	xenTro2	MAF Net
Zebra finch	<i>Taeniopygia guttata</i>	Jul. 2008	taeGut1	Syntenic Net
Zebrafish	<i>Danio rerio</i>	July 2007	danRer5	MAF Net

## APPENDIX II

### BigFoot: Additional Information

#### *The Alignment Transducer*

In general, the alignment transducer of BigFoot models indels in a similar way as the TKF92 statistical alignment model, and is placed on each branch of the phylogenetic tree in order to model the evolution from each ancestor to each descendent. Transitions to a slow state only occur when there is a ‘slow’ character emitted in the ancestral sequences. The same is true for transitions to a fast state with ‘fast’ characters. The model switches between fast and slow states only when a character type switch is emitted from the ancestral root, based on a basic hidden Markov model.

The transducer model contains two ‘wait’ states, W1 and W2, which is used while waiting for input from the ancestral sequences. The transducer will remain in the wait state until the next input character is emitted. For the second wait state, only the delete state can access it as delete state cannot self-transition. A complete explanation of the model can be found in BigFoot’s original paper by Satija *et al* (in press).

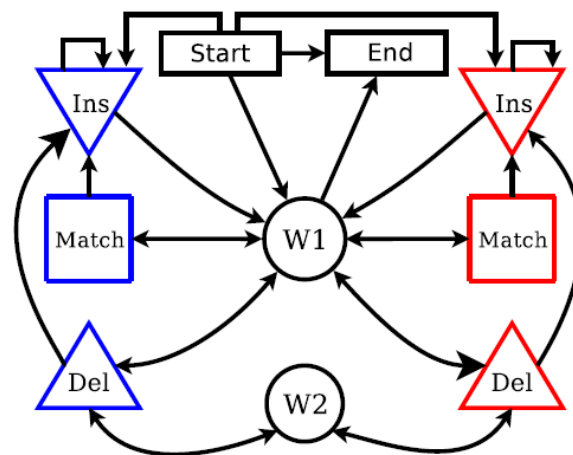


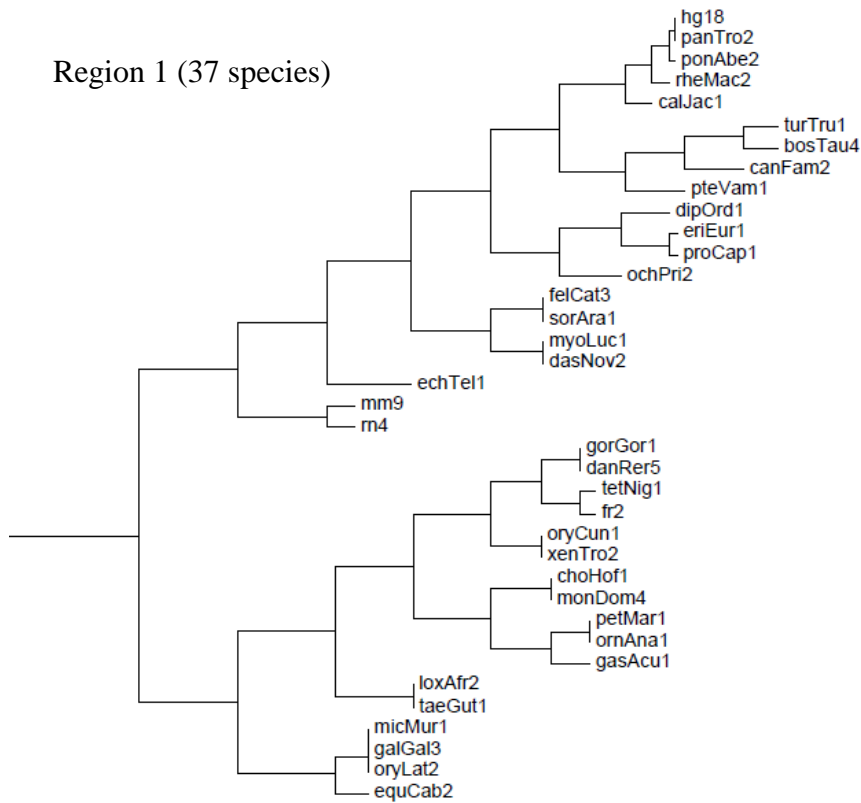
Figure above is the alignment transducer model. Match, Insert (Ins), and Delete (Del) states represent the evolutionary events. States with blue outlines are the slow states, which mean they have reduced evolution rates compared to states with red outlines, which are the fast states. (Image from Satija *et al*, in press)

## APPENDIX III

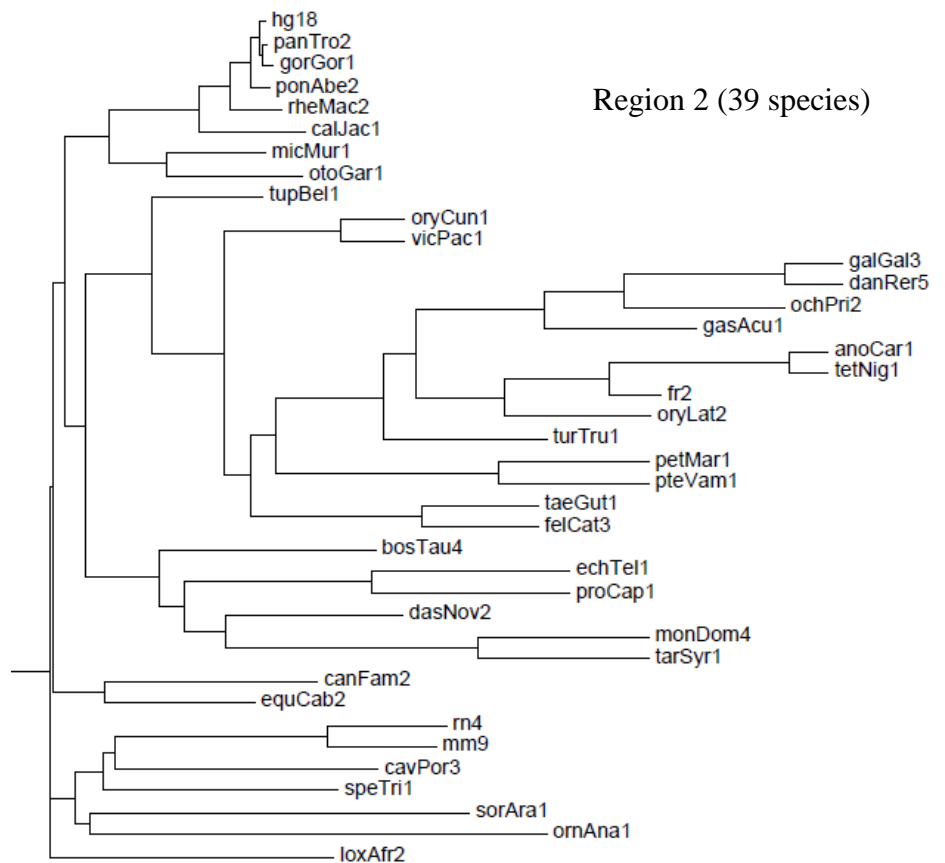
### Phylogenetic trees

*Phylogenetic trees produced by each set of species obtained from cons44way track in UCSC Genome Browser.*

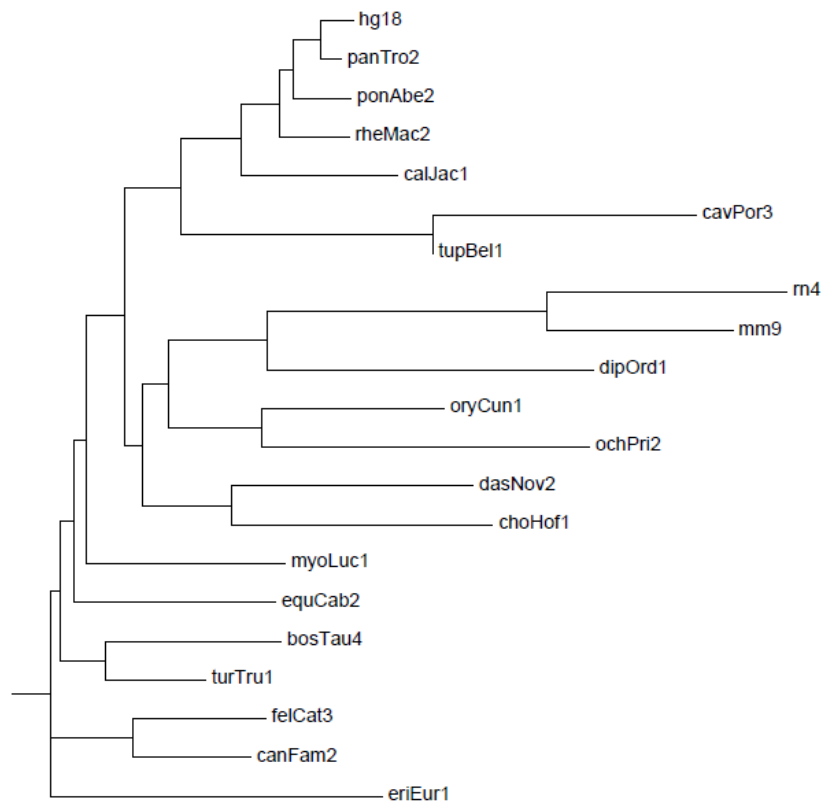
Region 1 (37 species)



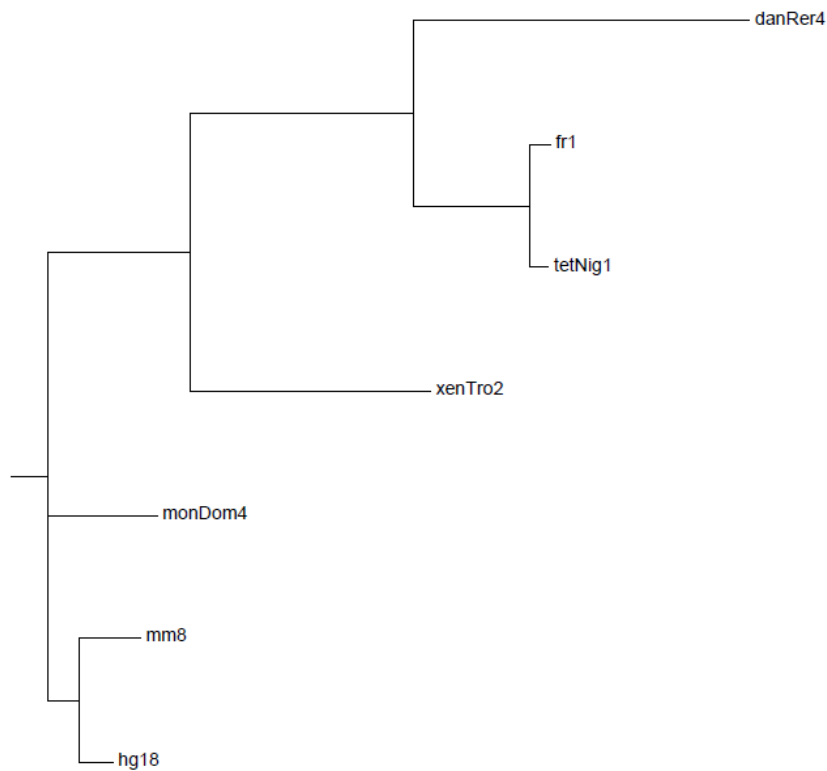
Region 2 (39 species)



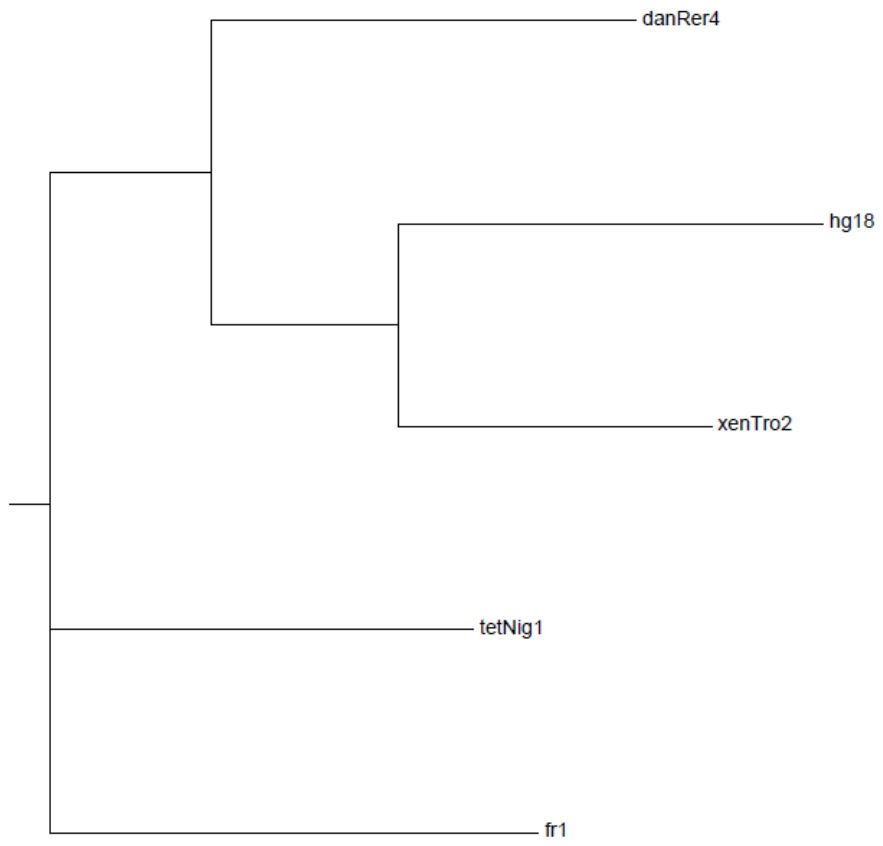
Region 3 (21 species)



Region 4 (7 species)



Region 5 (5 species)





### Region 3

hg18_3/1-352	1	t a c t c a t c a t c t t a a g c t c t c t c a a a t a t t t c c t t t t c t g g t c c t g c t g c o c t - t g c a t c t g - - - t t c a - - - - g a t a - - - g t g c t a t c a t c a c t t g a a c t c a t a t a t c t	103
felCat3_3/1-240	1	c a c c a a t t t c t c t c a a a t t c t c c c a a g a t t t c t c t t t c t f a t c t c t g t c t - c c t c c a g c t c t c a - - - g g t g - - - g t a t t c t c a t c a c t t g a a c t g - - - - a t t t	101
oruCun1_3/1-250	1	c a t c c a t c c c c t t a t g t t c t c t c a a a t t t t c c t t t c t g g t c c t a c t a c c c - c a c c c a g - - - c t c a - - - - g g t g - - - c t t g c a t g c a t c a c t t a a a c t g a c a c a g t c c	103
myoLuc1_3/1-220	1	- - - - - c a a t a c t t c c t t c t c t g a t c c t g t g c c c - c a c g a c a - - - g t a - - - - a g a g g a g t c a t t a t c a t c a c t t g a a c t a a t a c c a t c t	81
equCab2_3/1-251	1	t a t t c a t t c c t c t c a a a t t c t g c a c a t a c t t c c c t t t c t g a t c c t c t g c c c c a c c a c c - - - c a t a a t g a t g g t g - - - g t c a t t a t c a c c a c t t g a a c t a a t a c c a t c t	109
hg18_3/1-352	104	t c t c a t t g t a t g t t c t c t a c t t t g t t t t c t c a a t a a t a t t c a t a t g c a a g t g a c t c t - - g c t g a c c c t g c t c t g g c t t - c a c t g t g t c c c c a c c a t a g a a c t c t c t g t	214
felCat3_3/1-240	102	t c t c t t t g c c t g c t c c c t a c c t t g t t c a t c a a t a a t g g t c a t t t g c a a g c c a a c t c t c g t t g g a t c c t g c t c t g c c c c t c t g g c c c a t c t a t a a g a c t t c c t a c t	215
oruCun1_3/1-250	104	t c c c c t a t t g t t c c c a t c o c t g t t c a t c a a t a a a t t c t t g t a c a a g g t g t c t c t - - a c t g g c c t c t g c t c t g g c t g c t c t c t g t g c c c a a t a g a a c c t c t a c t	214
myoLuc1_3/1-220	82	t c - - - t g t t g t c t c t c a c a t t g t t c a t c a a t a g t a t t c a t t t g c a a g t a - - - t - - - g t t g a a c c t g c t c t g g c c t - c t c t g t c c t a c c a g c a a c c t t c t a c t	184
equCab2_3/1-251	110	t c c c c t g - - t g t t c c c t g c c t t g c t c a t c a a t a a t t c a t t t c a a g t g a t t c t - - g t t g a c c c t g c t c t g g c c t - c t c t g t c c c a t c a a t a a g c g c t c c t g c t	218
hg18_3/1-352	215	G C C T C c a g t t a c a t t t c t a a g t - - - - - t c - - - - - a a a t c t g a c t c c t g g g c t g g g c g g t g g c t a t g c c t g t a a t t c a g a c a c t t t g g g a g g c c a a t g t g a g	309
felCat3_3/1-240	216	G C T T C c a g t a a t a t t t c t a a g t - - - - - t c - - - - -	240
oruCun1_3/1-250	215	G C T T C c a g t a t g t t t c t c a g a t g g a g c g a c c c - - - - -	250
myoLuc1_3/1-220	185	G C T T C c a g t a t g t t t c t g a a g t - - - - - t c g a t c t g a - - - - -	217
equCab2_3/1-251	219	G C T T C c a g t a t a c t t c t c a a g t - - - - - t c a g a t c t g a - - - - -	251
hg18_3/1-352	310	t g g a t c a c t t g a g g t c a a g a t t c a a g a t c a g c a g g c c a a c a - - - - -	352
felCat3_3/1-240		- - - - -	
oruCun1_3/1-250		- - - - -	
myoLuc1_3/1-220	218	- - - - - c c t	220
equCab2_3/1-251		- - - - -	

### Region 4

denRer4_4/1-1001	1	t g g c a g - A A T G G T G C T G A g g a g a a g t g t c c c t g t t a a a t a g t t g g - - - a g a g c c a g t g c g - - - - - c c a t g t g a c g g - T T A C	74
fr2_4/1-268	1	c a g c g a - A G T G G T G C T G A a g g c g c a c a g g g c c t c t t a a a t a g c c g t t g g a c t c g g t t c g - - - - - g c a t g t g a c g g - T T A C	77
tetNig1_4/1-268	1	c a g c g a - A G T G G T G C T G A a g g c g c a c a g g g c c t c t t a a a t a g c c g t t g g a c t c g g t t c g - - - - - g c a t g t g a c g g - T T A C	77
mm9_4/1-311	1	c a g c g c T A G T G G T G C T G A a - - - - - c a a t a c t t c c t t c t g a t c c t g t g c c c - c a c g a c a - - - g t a - - - - a g a g g a g t c a t t a t c a t c a c t t g a a c t a a t a c c a t c t	69
hg18_4/1-310	1	c a g c g c T A G T G G T G C T G A a - - - - - a g a g c c g g c g c g c g c t t a a a t a g g a c t a t g c c a t g t g a t g g a c t a c	69
denRer4_4/1-1001	75	G C t a c t A G C G g c t g a g a g c t C T C A A T A G G G T T T - T G T G C C T T T - T C T C T G G C A T T c a c t C T G C A C G C T T C C C A T T A T T T C A C T T G T T A C T A A A T G A A C T G C A T A A T G T A	184
fr2_4/1-268	78	G C t g t a A G C G g c c a a a a g c t C T C A A T A G G G T T T - T G T G C C T - T - T C T C T G G C A T T c c c a C T G C A C G C T T C C C C A T T A T T C A C T T G T T A C T A A A T G A A C T G C A T A A T G T A	186
tetNig1_4/1-268	78	G C t g a A G C G g c c a a a a g c t C T C A A T A G G G T T T - T G T G C C T - T - T C T C T G G C A T T c c c a C T G C A C G C T T C C C C A T T A T T T C A C T T G T T A C T A A A T G A A C T G C A T A A T G T A	186
mm9_4/1-311	70	G C c a g g A G C G g c c g g g c c c t C T C A A T A G G G T T T T G T G C C T - T T C T C T T G G C A T T c a c t C T G C A C G C T T C C T A T T A T T T C A C T T G T T A C T A A A T G A A C T G C A T A A T G T A	180
hg18_4/1-310	70	G C c a g g A G C G g c c g g g c c c t C T C A A T A G G G T T T T G T G C C T - T T C T C T T G G C A T T c a c t C T G C A C G C T T C C T A T T A T T T C A C T T G T T A C T A A A T G A A C T G C A T A A T G T A	180
denRer4_4/1-1001	185	G T C T A T T C T T G T C A G C C C C A C C C A C a c t g c a a g a g g c g g t a c T G C - C C - - - - - c t t c c c t t t t a a a	246
fr2_4/1-268	187	G T C T A T T C T T G T C A G C C C C A C C C A C g c c g c a a g a g g c g g c a c T G C - C C - - - - - a t t c t c t c t t t - - -	245
tetNig1_4/1-268	187	G T C T A T T C T T G T C A G C C C C A C C C A C g c c g c a a g a g a c g g c a c T G C - C C - - - - - a t t c t c t c t t t - - -	245
mm9_4/1-311	181	C T C T A T T C T T G T C A C C C C A C C C A C g c c g t a a c a g g c g c a c t C G C C C c c c t c t c c c t t t t c g g g t t a t t t t c a t t c t c c c g a t t t a a t a t t c t t t c t t - - -	289
hg18_4/1-310	181	C T C T A T T C T T G T C A C C C C A C C C A C g c c g t a a c a g g c g c c c t C G C - C C c c t c c t c c c t t t t c g g g t t a t t t t c a t t c t c c c g a t t t a a t a t t c t t t c c t t - - -	288
denRer4_4/1-1001	247	a t t t c c c c t c c t t t t a t a c c a t a c t c t t t a g t g t c g t t t a a a c t t a t t t c t T T C T a c g c g a t c a c a t g g a c a c t a t g c t c t t t a a t t a t t c a a t t c a a g a g g g	358
fr2_4/1-268	246	- - - - - a t c a t c a t T C C A t t t a t c a t a - - - - -	268
tetNig1_4/1-268	246	- - - - - a t c a t c a t T C C A t t t a t c a t a - - - - -	268
mm9_4/1-311	290	- - - - - c t c t t a T T - T C t t a c t t t c c t t - - - - -	311
hg18_4/1-310	289	- - - - - c t c t t a T T - T C t t a c t c c c t t - - - - -	310
denRer4_4/1-1001	359	c c g t c g t t t g t t c c g c g t g a a c t c a t t a g a c a t t c g g t g c a t a c g c g t a c g t c a g c a g g c t c a t c g g c c c g g t t c a c c t g c g a c t c a g c c a t c a t t a a c t a t t a t a a c a	470
fr2_4/1-268		- - - - -	
tetNig1_4/1-268		- - - - -	
mm9_4/1-311		- - - - -	
hg18_4/1-310		- - - - -	
denRer4_4/1-1001	471	a t t a c a g a c c t t g t t t c c a g a g c g t g t t c g g g a g t a t g a a a a t a a a t a a t a a g a t c g a t t a t t g c a g t c g t t t g t g t a c g c c a c t g a c g t g c g t t a g t t g c g c a	582
fr2_4/1-268		- - - - -	
tetNig1_4/1-268		- - - - -	
mm9_4/1-311		- - - - -	
hg18_4/1-310		- - - - -	
denRer4_4/1-1001	583	t g g a a c a t c t g c t g a g a t c t t t c t t c c g a a c t a a c a t t t a a t t c g g a t a c c g a t c t g t c t c a t t t a g g a c g a t t t t t c a a a t t t a a t t a t g a t t a t t t t t t c c	694
fr2_4/1-268		- - - - -	
tetNig1_4/1-268		- - - - -	
mm9_4/1-311		- - - - -	
hg18_4/1-310		- - - - -	
denRer4_4/1-1001	695	t t a a g g a c a t t a t c c a t a g c t a g t c a a c a t c t g a a g t g g c a a c g t a t t t a a a a t c g t t t t a c a c t a a t a a a a a g g a t a t g a c a a a a g t t a a g a c a a c a g a t a t t	806
fr2_4/1-268		- - - - -	
tetNig1_4/1-268		- - - - -	
mm9_4/1-311		- - - - -	
hg18_4/1-310		- - - - -	
denRer4_4/1-1001	807	t g c a t g t g c t t t a a c t a t t t t a a c a g t c t a t c a g g t t c a t c t a c a t t t t t a a g t t a t a a c c t a a t a g t t c a g t c a g t t g a t t g a t g a a g t t g a g a t g a c t c g a a a a t	918
fr2_4/1-268		- - - - -	
tetNig1_4/1-268		- - - - -	
mm9_4/1-311		- - - - -	
hg18_4/1-310		- - - - -	
denRer4_4/1-1001	919	t t a a g g c a g c a a c a t c t t t t a c a g t g a a t g t g t t c a a t c g t t t c a g g a c a a c t t t g a t g a a a g g t t t c a t c c a c t t c a a a t	1001
fr2_4/1-268		- - - - -	
tetNig1_4/1-268		- - - - -	
mm9_4/1-311		- - - - -	
hg18_4/1-310		- - - - -	

### Region 5

denRer4_5/1-229	1	g t g a c t t t A A A T T C G C C T C G A A G C T G A C T T - - - - - C T T T T G T C T A a g a t t t a a a t g c g g a t t g t g c t g t t g t A T T A A A A T A C A T T C A A T T A A A A T T G G A A T T T	105
hg18_5/1-115	1	a t - - - C T T C T A C T C A C C A T T C C C C A C T G G C T T - - - - - C C T T G T G C - - - t t c c t t t t a c t c t c a a t a c t t c a t g c a t t a a g a a g c g t a a a a t t t a a a a g t t g a c t t t	100
tetNig1_5/1-198	1	g g a a c A T T G C T T G A G C G A C C C A G C A T G A T T G G G G T C C T T G G C A - - - g t t a t a a g c c t c a t a a t t a g g g a a c a a a c c g c t t a g a g t a t t c c a t t c t g c a a t t t	108
fr2_5/1-190	1	g g a g c C T T T A T T T A C T C C T G A G G A T G A A C T - - - - - C A T G T G G C A - - - c a a c a g c t c g g t t a c g c a c g g a c g c t t g T C T G C C T A T A G T C T G C G T A C A G C T T G A A C G T T	103
xenTro2_5/1-132	1	t t a a a A T T A G T A A A T T A A G G T C T G G G A T T T G - - - - - T A T T T T T G - - - a t t c a t t g c a t c c c - - - - - t A G T T G C C A A T A A C A T A C A T C A A G A T G T G C T G T	89
denRer4_5/1-229	106	T C a t t g g g c a c a t t t t a c g a t a t t t t a a c t c a a g t g c t t t a c g g t t a t a a a g a t t a g a t g a g t c a c t a t a t g t t g g a a c a t t a t t a a t t t t c a t t t t a t	215
hg18_5/1-115	101	C C a c c t g a g g t a t t t - - - - -	115
tetNig1_5/1-198	109	T A g a t g c a a a t g a a c t t t g a c g c t c g c a a c a g t c g g c a c - - - - - a a c c a t a a g t a g a c t a a g t a t a a a a t g t c t c g t a t	184
fr2_5/1-190	104	- - - G T a t g a t g a c c c c c a a t a t c a t t t t a g c t c t t t g g - - - - - c a c a a a c t t a g a a c c t a t t c a c g t c t g a t t a t g	176
xenTro2_5/1-132	90	C - - - - - c c a c t a a c a a g g a t a a c c a g g a t a a c c a g g c t g	118
denRer4_5/1-229	216	t t a a c c a t a t g c c t a - - - - -	229
hg18_5/1-115		- - - - -	
tetNig1_5/1-198	185	g t t g a g g t t a t g t t - - - - -	198
fr2_5/1-190	177	g a g g c a a a t g c t t a - - - - -	190
xenTro2_5/1-132	119	g t g c a c t t t a t t t - - - - -	132