TRACING THE EVOLUTION OF LINEAGE-SPECIFIC TRANSCRIPTION FACTOR BINDING SITES IN A BIRTH-DEATH FRAMEWORK

BY

YANG ZHANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioengineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Associate Professor Jian Ma

ABSTRACT

Changes in cis-regulatory element composition that result in novel patterns of gene expression are thought to be a major contributor to the evolution of lineage-specific traits. Although transcription factor binding events show substantial variation across species, most computational approaches to study regulatory elements focus primarily upon highly conserved sites, and rely heavily upon multiple sequence alignments. However, sequence conservation based approaches have limited ability to detect lineage-specific elements that could contribute to species-specific traits. In this thesis, we describe a novel framework that utilizes a birth-death model to trace the evolution of lineage-specific binding sites without relying on detailed base-by-base cross-species alignments. Our model was applied to analyze the evolution of binding sites based on the ChIP-seq data for six transcription factors (GATA1, SOX2, CTCF, MYC, MAX, ETS1) along the lineage toward human after human-mouse common ancestor. We estimate that a substantial fraction of binding sites (58-79% for each factor) in humans have origins since the divergence with mouse. Over 15% of all binding sites are unique to hominids. Such elements are often enriched near genes associated with neural-related functions and pathways, and harbor more common SNPs than older binding sites in the human genome. These results support the ability of our method to identify lineage-specific regulatory elements and help understand their roles in shaping variation in gene regulation across species.

To my parents, for their love and support.

ACKNOWLEDGMENTS

First and foremost I would like to express my deepest thanks to my advisor, Jian Ma, for his great guidance on this project, inspiration, motivation and valuable advice throughout the past few years. I appreciate all his contribution of time, ideas and funding to make my research productive and stimulating. I am also thankful for his agreement to accept me as a member of Ma lab. He showed great patience and gave my full support to allow me find my research topic and continue my PhD study.

I would also like to thank my colleague Ken Yokoyama who is also the co-author of this project. I benefited a lot from discussions with him and it is this study lead to my current research topic and I enjoy it very much.

Members of Ma group are always very supportive and have contributed immensely to my personal and professional time at UIUC. I would like to thank Yang Li, Shuomeng Guang, Jack P. Hou, Yiyi Liu and Brittany Weida for their patience and stimulating discussions. I would also like to thank other friends in IGB: Xinlei Wang, Jichuan Zhang, Jingyi Fei and Xuefeng Wang.

Last, I would like to thank my family: my mother Jinyan Wang and my father Donghe Zhang, for giving birth to me at first and for their their love and encouragement throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION[1]

Changes in gene regulation play a key role in the evolution of morphological traits (Wray, 2007; Davidson, 2001; King and Wilson, 1975). At the level of transcription, gene expression is controlled via transcription factor (TF) proteins that selectively bind to cis-regulatory elements in a sequence-specific manner (Davidson, 2001; Wray et al., 2003). Utilizing chromatin immunoprecipitation of specific TFs followed by high-throughput sequencing (ChIP-seq), recent studies showed that the evolution of these transcription factor binding sites (TFBS) is highly dynamic, with sites differing a great deal even within mammals (Odom et al., 2007; Schmidt et al., 2010; Bourque et al., 2008; Scally et al., 2012; Borneman et al., 2007).

Despite substantial experimental evidence for rapid divergence of regulatory protein-binding events across species, computational models designed to analyze regulatory elements using cross-species comparisons have focused primarily upon phylogenetic footprinting approaches, in which putatively functional regulatory elements are identified according to sequence conservation (Kasowski et al., 2010; Siepel et al., 2005; Hardison et al., 1997; Blanchette et al., 2006; Kellis et al., 2003; Margulies et al., 2003). Previous computational studies have inferred the evolution of regulatory elements using, for example, the emergence of new conserved elements specific to a particular clade in the phylogeny (Lindblad-Toh et al., 2011) or lineage-specific alterations leading to a loss-of-function phenotype (Lowe et al., 2011; Hiller et al., 2012). Although such approaches have been helpful in understanding lineage-specific regulatory element evolution, all inherently rely upon fixed cross-species alignments, which are frequently of low quality within

[1]This study previously appeared as an article in the Journal of *PLoS Computation Biology*. The original citations is as follows: Yokoyama KD†, Zhang Y† and Ma J. Tracing the Evolution of Lineage-Specific Transcription Factor Binding Sites in a Birth-Death Framework. *PLoS Computational Biology*, 10(8):e1003771, 2014 (†: equally contributed Author)

non-coding regions in the genome (McLean et al., 2011; Chen and Tompa, 2010; Kim and Ma, 2011). Previous studies have estimated that more than 15% of aligned bases within human-mouse whole-genome alignments are incorrect (Majoros and Ohler, 2010) and the error rate increases when more species are involved (McLean et al., 2011). Ancestral reconstruction, which is sensitive to details of the multiple alignment, is a particularly challenging problem for non-coding regions (He et al., 2009; Lunter et al., 2007). As a consequence, cross-species comparisons of non-coding sequences are limited in their ability to study regulatory sequence evolution, particularly in cases where the elements are selected for novelty or newly-derived. Such newly-derived regulatory elements are not rare; indeed, analyses using human population variation data from the 1000 Genomes Project (Margulies et al., 2003) have shown that human genomic locations under selection undergo considerable turnover and frequently lie outside mammalian-conserved regions (Blanchette et al., 2004). Yet, systematic identification of binding sites for specific TFs and assessment of their conservation and prevalence using cross-species comparisons remains a challenging problem.

In this work, we introduce a novel evolutionary framework through which lineage-specific TFBSs can be inferred on a genome-wide scale. In contrast to conservation-based approaches (Blanchette et al., 2006; Lindblad-Toh et al., 2011; 1000 Genomes Project Consortium et al., 2010), we utilize a birth-death model to infer ancestral states of a given motif without the use of the base-by-base alignment details in the underlying cross-species sequence alignment (details in Chapter 2). Gains and losses of TFBS have been explicitly used both to improve cross-species sequence comparisons and to detect cis-regulatory modules, although such models are usually framed within the context of an alignment (Kim and Ma, 2011; Ward and Kellis, 2012). A more similar alignment-free model was previously used to measure the overall rate of TFBS creation along different lineages (Xie et al., 2005). In this work, we instead applied our framework to infer lineage-specific TFBS, estimating the branch of origin of each individual TFBS for six TFs. Chapter 3 presents results about studying patterns for TFBS with different branches of origin, including target genes of the newly-derived sites and relationship with within-human variation. Our results provide strong support that this novel method can help identify lineage-specific regulatory elements, a first step towards understanding the role of regulatory element evolution in shaping the

variation of gene regulation across species. Finally in Chapter 4 we provide a brief summary of this thesis, its limitations and directions of future work.

CHAPTER 2

METHOD

2.1 Method Overview

Our goal is to detect lineage-specific rates of TFBS evolution and the branch of origin for individual TFBS. Here, lineage means any ancestral branches in the phylogeny or a branch leading toward any modern species. Our approach is to model TFBS evolution using a birth-death framework, in which individual TFBSs can be gained, lost, or conserved within a given lineage during evolution. The rate of TFBS creation (named birth rate) and loss (named death rate) are first estimated from a set of orthologous sequences, and are subsequently used to trace the evolutionary origin of individual TFBSs at the sequence level. The birth rate (α) for a given motif represents the probability at which a TFBS appears at a single unoccupied site in a given year of evolutionary time. Similarly, the death rate (β) represents the rate at which an existing TFBS is lost per year. The method considers only TF motif counts within orthologous sequences across species, and therefore does not require an accurate base-to-base multiple sequence alignment. This framework allows us to reconstruct the ancestral states for each TFBS throughout the genome, providing a distribution for the branch of origin of the binding sites genome-wide. The main scheme of our method is shown in Figure 2.1.

For any set of orthologous sequences across species and a known phylogeny, we first estimate birth and death rates according to the observed numbers of TF motif occurrences within each species. Such orthologous sequences can, for instance, be obtained using a genome-wide multiple species alignment (MSA). However, the underlying base-level alignment is ignored once the orthologous sequences are obtained, and subsequently the model considers only the number of TF motifs within each sequence. Thus, the method operates independently of any details within the alignment once the sequence

correspondence between species (i.e., orthologs) is obtained. Every node in the phylogeny is then associated with a (random) variable Q_x, which represents the number of occurrences of the TFBSs at that node. The value of Q_x is known for each leaf node in the tree for any given ortholog set. Birth and death rates of a given motif can then be estimated by maximizing the likelihood across the entire data set, taking into account both branch lengths as well as the size of the sequence region (see Section 2.3 for mathematics details). Evolutionary rates were estimated using an iterative approach, but were found to be extremely robust according to the initial parameter settings. Once the birth and death rates are determined, we can use these rates to trace the branch of origin of individual TFBSs. This can be done by firstly reconstructing the most likely ancestral state at each node of the phylogeny; i.e., the value of Q_x that maximizes the likelihood of the data for each individual ChIP-seq peak region. This provides the most likely number of TF motif occurrences at each node, and allows us to trace the most likely branch of origin for individual site. The overall procedure of our method works as follows.

1. We identify motif occurrences under human ChIP-seq peak region and its orthologous sequences within the MSA block.

2. We estimate and maximize the likelihood for each ancestral node in the phylogeny given the motif occurrences in the descendant species in a iterative way.

3. We determine the branch of origin for the TF-bound motif within ChIP-seq peak regions.

2.2 Determine motif position in extant species

In this study, we restrict motif search within human ChIP-seq peak region. ChIP-seq data set can be got from public sources such as ENCODE (ENCODE Project Consortium, 2011). We then obtained orthologous sequences across 46 vertebrate species genome-wide corresponding to ChIP-seq peaks in humans using the 46-way multiz alignments (Miller et al., 2007) available at the UCSC Genome Browser (Karolchik et al., 2004). To determine the

binding motif of each TF and to determine the branch of origin for each TFBS, we used the (−100,+100) region window relative to the summit of the peak in humans.

TF binding motifs were predicted by clustering 7-mers with increased branch-specific evolutionary rates using the likelihood ratio test as shown in Section 2.3.5. Since the minimum length of most motifs in the JASPAR Core Database is approximately 7bp (Sandelin et al., 2004), for each TF we tested for increased lineage-specific evolutionary rates across 7-mer seeds. To circumvent an exhaustive search across the comprehensive list of all 7-mer motifs, we limited our motif scan to the 1500 most frequently occurring 7-mers in each data set, which was generally the number of 7-mers exhibiting statistically significant over representation within the sequences. We then iteratively clustered significant 7-mers (P-value $< 1e^{-10}$ according to likelihood ratio test) that are predicted along a branch ancestral to humans. At each step, a 7-mer was placed into an existing cluster if it was found to similar to another 7-mer in the cluster with at most one mismatch. As many 7-mers were predicted under the threshold, we filtered redundant clusters containing an identical 7-mer, keeping the cluster producing the highest P-value. Each cluster thus comprised a set of aligned 7-mers, which was condensed into a single consensus sequence using criteria similar to (Matys et al., 2003; Cavener, 1987), where each column was assigned single residue if it comprised at least 50% of the total score and at least twice the score of every other nucleotide. In the remaining cases, double nucleotide degeneracy was assigned to sites in which at least 75% of the total score was attributed to two nucleotides, otherwise the site was considered fully-degenerate.

From the consensus sequences generated from the previous step, PWMs were generated using an iterative approach in which the initial consensus sequence was converted to a PWM. This generates a list of k-mer motifs, and motifs are allowed to be on either strand. Based on this list, an initial PWM can be constructed using the observed data within the (−100,+100) window surrounding all ChIP-seq peaks. A new cutoff is then set, using a score:

$$S = \log\left(\frac{\prod_{i=1}^{k} c_j(w_i)}{\prod_{i=1}^{k} b(w_i)}\right) \tag{2.1}$$

where $c_j(w_j)$ is the frequency of nucleotide w_i in a given k-mer $w = w_1 w_2 ... w_k$ that is a potential binding site, and $b(w_i)$ is the background frequency of

nucleotide w_i.

This score represents the quality of the match of the potential k-mer over the expected match. The threshold is set so that, with each iteration, the score of the lowest-ranking k-mer is equal-to or above that of the new motif. In other words, we want to include a large number of peaks using a small number of k-mers. PWMs are updated after each iteration using this approach, until the list of k-mers no longer changes between successive iterations. The final PWM is then used to generate a final list of k-mers. The threshold is set to the score at which the number of peaks containing a binding site increases less than expected from a random list of k-mers, and also includes at least 60% of the total number of peaks.

Finally, we scan for motifs in orthologous regions in extant species within $(-100,+100)$ window centered on human ChIP-seq peaks. The number of motif occurrences in extant species was used as the input of our birth-death model.

2.3 Birth-death framework

The foundation of our approach to estimate genome-wide rates of evolution for a given motif is based on a birth-death framework (formally, a quasi birth-death process (Cavender, 1978)), similar to that used to measure the timing of accelerated motif evolution as in (Yokoyama and Pollock, 2012). In our modal, we assume the evolution of TFBS can be modeled as a combination of birth and death events which are determined by two parameters. The birth rate (α) represents the rate at which a new motif occurrence appears at any unoccupied site per year, while the death rate (β) represents the rate at which an existing site is lost per year. Given a set of orthologous sequences and a known phylogeny, we estimate birth and death rates for the motif across the phylogenetic tree using a maximum likelihood approach. The following sections will explain each step in details.

2.3.1 Determining the probability of TFBS turn over

Suppose $w(t)$ is the probability that a TFBS exists in time t. The probability of observing binding site at same position in time $t + 1$ under birth-death

model is then:

$$w(t+1) = \alpha(1 - w(t)) + (1 - \beta)w(t) \tag{2.2}$$

The rate of change of probability $w(t)$ over time t can be written as $w'(t) = w(t+1) - w(t)$, which gives Equation 2.3 by taking the differential equation of Equation 2.2:

$$w'(t) = \alpha - (\alpha + \beta)w(t) \tag{2.3}$$

Solving differential Equation 2.3 gives two solutions depend on whether there is a binding site at time $t = 0$. We denote those two solutions by $u(t)$ and $v(t)$, where $u(t)$ assumes that the motif was present at this site at time $t = 0$ ($w(0) = 1$), while $v(t)$ assumes that the motif did not exist at time $t = 0$ ($w(0) = 0$). It is noted that both $u(t)$ and $v(t)$ are solutions for Equation 2.3, both represent the probability that the motif exists at a specific position after time t, differing only in the initial conditions. So the solutions of Equation 2.3 give:

$$u(t) = \frac{1}{\alpha + \beta}\left[\alpha + \beta e^{-(\alpha+\beta)t}\right] \tag{2.4}$$

and

$$v(t) = \frac{\alpha}{\alpha + \beta}\left[1 - e^{-(\alpha+\beta)t}\right] \tag{2.5}$$

In a region with N nucleotides, suppose there are i occupied binding sites at time $t = 0$, then the probability that k binding sites remain after time t can be derived from probability mass function of binomial distribution:

$$U_{i,k}(t) = \frac{i!}{(i-k)!k!}u(t)^k(1 - u(t))^{i-k}, k \in \left[0, i\right] \tag{2.6}$$

Similarly, the probability that there are b binding sites generated from $N - i$ unoccupied nucleotide sites will be:

$$V_{N-i,b}(t) = \frac{(N-i)!}{(N-i-b)!b!}v(t)^b(1 - v(t))^{N-i-b}, b \in \left[0, N - i\right] \tag{2.7}$$

The transition probability $p_{ij}(t)$ represents the probability that the given

region have j binding sites after time t with initial binding site number i:

$$p_{ij}(t) = \sum_{k=0)}^{min(i,j)} U_{i,k}(t) \cdot V_{N-i,j-k}(t) \tag{2.8}$$

Here, the sum is over all possible values k, where k represents the number of motif occurrences at time t among the sites that were originally occupied at time $t = 0$

2.3.2 Calculating the likelihood of the data

Given the birth and death rates (α and β) across the tree (which are estimated using the method described in Section 2.3.4), we can calculate the likelihood of the data using Felsensteins pruning algorithm (Felsenstein, 1973). Let us first consider data from a single sequence. We let $\theta = [\alpha, \beta]$ represent the paramter vector comprising the birth and death rates, and let D_Y represent the data downstream of a node Y in the phylogeny. Let $Z_1, Z_2, ..., Z_m$ be the daughter nodes of Y, occurring at times $t_{Z_1}, t_{Z_2}, ...t_{Z_m}$ relative to parent node Y, respectively.

If random variable Q_y represents the number of motif occurrences at node Y, the likelihood $x_Y(i;\theta) = \Pr(D_Y \mid Q_y = i;\theta)$ of the data downstream of Y, assuming i motif occurrences exist at node Y can be obtained recursively. This likelihood is given by:

$$x_Y(i;\theta) = \prod_{k=1}^{m} \sum_j p_{ij}(t_{Z_k}) \cdot x_{Z_k}(j;\theta), \tag{2.9}$$

where the inner sum is across all possible values for j, corresponding to the number of motif occurrences at daughter node Z_k. If node Z is an extant lineage, the probability $x_Z(j;\theta)$ is equal to 1 if we actually observe j motif occurrences within the sequence, otherwise the likelihood is zero.

The likelihood of the data can therefore be obtained recursively by determining the values $x(i;\theta)$ progressively for each node farther up the tree. The log-likelihood $L(D^{(k)};\theta)$ for the k-th sequence is then given by:

$$L(D^{(k)},\theta) = \log\big[\sum_j P(j) \cdot x_R(j;\theta)\big], \tag{2.10}$$

where R is the root node and $P(j)$ is the prior probability that j binding sites exist in a single sequence. For our implementation, prior probabilities $P(j)$ was set to the Poisson distribution: $P(j) = \lambda_j e_{-\lambda}/j!$ where λ is the mean number of motif occurrences per sequence. The total log-likelihood $L(D:\theta)$ is then the sum $L(D;\theta) = \sum_{k=1}^{n} L(D^{(k)};\theta)$ across each of the n regions.

2.3.3 Determining the optimal ancestral states

We can determine the most likely ancestral states using the computed values for $x(j;\theta)$ at each node in the phylogeny. At the root node R, the most likely ancestral state is the one that produces the highest likelihood; that is, the value of j that maximizes the expression $q_R = \text{argmax}_j P(j) \cdot x_R(j;\theta)$. Progressively moving down the tree, if the most likely number of motif occurrences at parent node Y is q_Y, then the optimal number of motif occurrences q_Z at a daughter node Z is given by:

$$q_Z = \text{argmax}_j x_Z(j;\theta) \cdot p_{q_Y j}(t_Z), \tag{2.11}$$

where t_Z is the time from node Y to node Z.

2.3.4 Birth-death rate estimation

Birth and death rates can be estimated using a maximum-likelihood approach. Namely, we use an EM-based approach (Dempster et al., 1977) to iteratively optimize the likelihood of the data D given the paramters $\theta = [\alpha, \beta]$. We begin with an initial estimate $\theta^{(0)}$ for the birth-death rates, generated by determining empirical birth-death rates after conducting ancestral reconstruction using parsimony (in our analysis, we found that the optimized parameters were not sensitive to the initial estimates). We determine the most likely ancestral state at each node given the initial parameter values. We then determine the observed number of births and deaths according to these optimal ancestral states, providing new estimates for the birth and death rates $\theta^{(1)} = [\alpha^{(1)}, \beta^{(1)}]$. We then continue the process, using the previous parameter estimates $\theta^{(i)}$ at each iteration to estimate the optimal ancestral states and obtain more optimal estimates of the birth and death rates $\theta^{(i+1)}$ until convergence (i.e., where $\left\|\theta^{(i+1)} - \theta^{(i)}\right\|^2$ falls below a certain

threshold).

2.3.5 Assessing branch-specific deviations in birth-death rates

To determine motifs or k-mers that exhibit branch-specific differences in birth-death rates along a specific lineage, we use a likelihood-ratio test. Namely, we compare the total log-likelihood $L(D;\theta_0)$ of the data according to the null model $\theta_0 = [\alpha, \beta]$, in which birth-death rates α and β are constant throughout the phylogeny, to the log-likelihood $L(D;\theta_A)$ of an alternative model $\theta_A = [\alpha, \beta, \alpha_A, \beta_A]$, in which birth-death rates α_A and β_A vary along a single branch relative to the rest of the phylogeny. For both models we estimate the parameters θ_0 and θ_A according to maximum-likelihood as described in aforementioned birth-death model. Our framework meets regularity conditions, and thus the scaled deviance $2[L(D;\theta_A) - L(D;\theta_0)]$ follows a chi-squared distribution with $|\theta_A| - |\theta_0|$ degrees of freedom (Davidson, 2001). P-values representing the statistical significance of lineage-specific acceleration along a specific branch can thus be determined using an F-test.

2.4 Method Evaluation

We evaluated the performance of our model in predicting the age of TFBS as compared to traditional methods based on phylogenetic footprinting approaches. To make a fair comparison, we first constructed a benchmark data set based on ChIP-seq data generated in human-mouse analogous cells used from ENCODE project. Positive cases (conserved TFBS originated before human-mouse common ancestor) are 200bp regions centered on human ChIP-seq peak summit that satisfy the following requirements: 1) For the same TF, there exists ChIP-seq peak in mouse analogous cells with mouse peak summit within +/-200bp of the human peak summit; and 2) TFBS exist in both human and mouse peak region. Negative cases (TFBS originated after human-mouse common ancestor) are 200bp regions centered on human ChIP-seq peak summits that do not have shared peaks in mouse and do not have TFBS in mouse orthologous region.

We compared our method with phylogenetic footprinting methods at both element level (using MotifMap (Xie et al., 2009)) and module level (using

PReMod (Blanchette et al., 2006)). Since phylogenetic footprinting approaches only predict whether a TFBS is conserved in a phylogeny but not provide information of specific lineage where the TFBS originated from, we conducted our evaluation by comparing the ancestral TFBS (i.e., the ones conserved between human and mouse) predicted by our method with the predictions from MotifMap and the predictions from PReMod. We used two windows sizes, ± 15bp and ±, 30bp, in order to directly compare with MotifMap (±15bp shift size). True positive (TP) means an ancestral TFBS (originated before human-mouse last common ancesotr) is predicted under positive benchmark data set defined above. False negative (FN) means no ancestral TFBS is found under positive benchmark data set. True negative (TN) and false negative (FN) are defined similarly using negative benchmark data set.

2.5 Figures and Tables

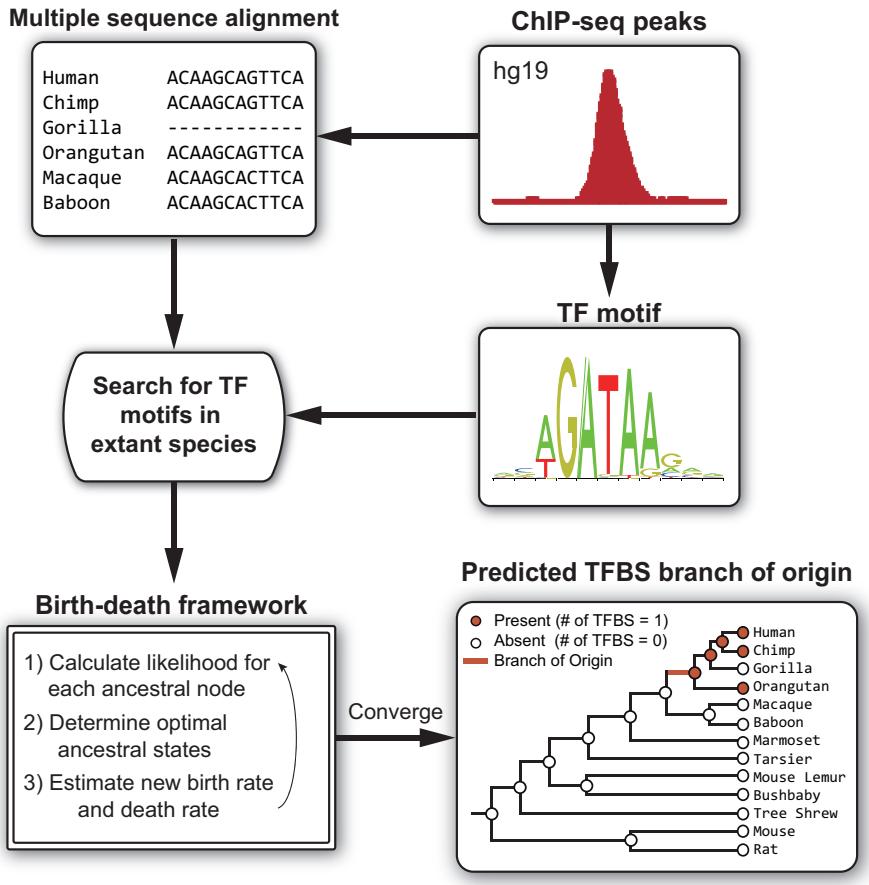

Figure 2.1: Overview of birth-death framework

CHAPTER 3

RESULTS

We applied our birth-death probability model to six TF ChIP-seq data, namely GATA1, SOX2, MYC, MAX, ETS1, and CTCF based on previous published data (Chen et al., 2008; Lister et al., 2009; ENCODE Project Consortium, 2011). Here, ChIP-seq data was used because it is now commonly used to map *in vivo* TF occupancy genome-wide (Johnson et al., 2007) and could reduce the false positive of searching TF occupancy using only motif PWMs. These TFs were chosen, in part, for their diverse functional attributes, their well-documented binding motifs, and the availability of ChIP-seq data in analogous cell types in human and mouse (see Section 2.4 for how we build benchmark data for evaluation).

Based on the phylogenetic tree we used from UCSC Genome browser, each binding site was thus either inferred to be present in the common human-mouse ancestor, or a more recent lineage leading to human. In order to assess the performance of our model, we first compared our predictions with other computational predictions made by footprinting based method and general conservation score such as PhyloP. It should be noted footprinting method usually only predict whether a TFBS is conserved in a phylogeny, while our model actually can also provide information of specific lineage where a TFBS originated from. Next, we further assessed our prediction using human and mouse ChIP-seq data from analogous cell type. Finally, we studied genomic functional patterns for TFBS associated with different branch of origin, including potential nearby target genes and relationship with human common variations.

The model framework and its motivations are illustrated in Figure 3.1. Figure 3.1(A) shows one scenario where the binding site was introduced to the genome through transposable elements (TEs) insertion followed by point mutation, which is most likely branch of origin of this site under our model. Figure 3.1(B) shows an example that our method is able to identify

cases of TFBS turnover within stationary modules that might not otherwise be detected using human-mouse ChIP-seq data direct comparisons. In this genomic region, there is a human GATA1 binding site originating on the ancestral primate lineage and a GATA1 binding site specific to mouse and rat. Although the ChIP-seq peaks appear in the same location between human and mouse, our model can predict such lineage-specific events (which is also reflected in the cross-species alignment). Again, our algorithm predicted these branches of origin accurately without base-by-base details of cross-species alignment.

3.1 Comparison with other methods

First, we compared the sequence level conservation of predicted TFBS according to their predicted branch of origin. We used the PhyloP mammalian conservation scores (Pollard et al., 2010) available at the UCSC Genome Browser to determine the conservation level for TFBS in human. For a specific TF, we first computed the average PhyloP score (X) in each ChIP-seq peak and then calculated the average score (M) as well as standard deviation (SD) across all peaks in the genome. We then grouped the binding sites according to their branch of origin (in four groups: Hominid-specific, Simian-specific, Primate-specific, and Eutherian-specific). Finally, we calculated the Z-score, i.e. $(X - M)/SD$) and compared Z-score distributions across four age groups. As expected, older binding regions show higher sequence level conservation than younger ones (Figure 3.2). These results suggest that our method can identify more recent, less-conserved TFBS, without relying on sequence-level conservation details.

Additionally, to further demonstrate the effectiveness of our method in identifying conserved TFBS, we directly compared with methods that use phylogenetic footprinting approaches. We compared with phylogenetic footprinting methods at both element level (using MotifMap (Xie et al., 2009) which is based on the method used in (Stark et al., 2007; Xie et al., 2007; Kheradpour et al., 2007)) and module level (using PReMod (Blanchette et al., 2006)). For the MotifMap method, we chose 1.91 or 40% BBLS score (equal to 60% confidence level according to (Stark et al., 2007; Xie et al., 2007)) as threshold to call a conserved TFBS. For PReMod method, ancestral regions

are defined as regulatory modules shared between human and mouse. We evaluated sensitivity (Equation 3.1), specificity (Equation 3.2), and accuracy (Equation 3.3) of these three methods.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{3.3}$$

The definition of TP, FN, TN, FP can be found in Section 2.4.

Overall, our method outperformed both MotifMap and PReMod in terms of accuracy (Table 3.1) . In addition to results from a fixed threshold for MotifMap used in the (Table 3.1) , we also plot the ROC curves when different threshold scores were used in MotifMap (Figure 3.3). For all TFs, our method outperformed MotifMap.

3.2 Substantial number of human TFBSs have recent origins in primates

Using our approach, we sought to determine the branch of origin for each human binding site for the six TFs. The distribution of the branch of origin for each TFBS is shown in Figure 3.4. Notably, between 58-79% of all human TFBSs had inferred origins after the human-mouse split.

To assess the accuracy of the age estimates, we compared our results to ChIP-seq data from human and mouse. Using analogous cell types across species, we determined the amount of overlap between human ChIP-seq peaks and ChIP-seq peaks in the orthologous regions in mouse. A human ChIP-seq peak was considered to be shared with mouse if its summit was within 200bp of a mouse ChIP-seq peak summit in the orthologous region (note that the mouse ChIP-seq data were not the input our method). The amount of overlap was assessed separately for regions containing a human binding site present in the common human-mouse ancestor and for regions that are not ancestral.

We emphasize that, as illustrated previously in Figure 3.1, ChIP-seq peaks shared across human and mouse can often contain TFBSs that are genuinely

lineage-specific, since ChIP-seq peaks span a relatively broad region and can contain instances of TFBS turnover within static modules. In addition, human-specific ChIP-seq peaks can also contain ancestral binding sites, since such sites can either be lost (non-conserved) along the mouse lineage or may not be bound by the TF along that lineage.

Table 3.2 shows the amount of overlap in ChIP-seq peaks between human and mouse according to the estimated branch of origin of the TFBSs. Human peaks containing predicted ancestral TFBSs were far more likely to overlap with bound regions in mouse than peaks containing only predicted lineage-specific sites. Between 24-41% of human peaks that overlapped with a peak for the same TF in mouse contained only predicted lineage-specific TFBSs, while 59-76% of shared peaks contained a predicted ancestral TFBS. Thus, there was a clear enrichment for TFBSs predicted to be ancestral among the ChIP-seq peaks shared between human and mouse. Among human-specific ChIP-seq peaks, a substantially greater number contained only lineage-specific TFBSs than sites predicted to be ancestral to human and mouse.

Although a relatively sizeable portion of shared ChIP-seq peaks contained only TFBSs predicted to be lineage-specific, in the majority of cases (>90%) the mouse TFBS did not occur within in sequence region orthologous to the human peak region used, but was instead offset to a non-overlapping region within a mouse peak. Very few of these TFBSs actually aligned across the two species, compared with those with predicted ancestral origin.

3.3 Within-species variation is higher among TFBSs of more recent origin

Recent work has reported a substantial difference between genomic locations that are conserved across species versus those conserved within the human population (Ward and Kellis, 2012). Thus, we compared human variation data to the relative age of the TFBSs. Comparing the overall frequency of common SNPs in humans among TFBSs originating at different times of evolution showed that a substantial fraction of human-specific TFBSs contained common SNPs, comprising over 6% of all human-specific TFBSs (Figure 3.5). This is much higher than the total fraction of TFBSs overlapping with

a common SNP, at a median of less than 3% across all six factors.

Since substantial variation exists in TF-binding events between human individuals (Kasowski et al., 2010), this high amount of variation among human-specific binding sites may partially reflect the fact that some TFBSs inferred to be human-specific may not be shared by the entire human population. However, recently-derived TFBSs in hominids were also substantially enriched for common SNPs, even when excluding human-specific TFBSs. For instance, among hominid-specific binding sites that are not human-specific, with a median of almost 4% of all sites. As these sites are shared across species, they cannot be fully explained by variation within the population. In contrast, common SNPs were consistently low among TFBSs with origins prior to hominids (Figure 3.5). Note that this observation was not biased by the SNP density surrounding the binding sites (Figure 3.6).

3.4 Hominid-specific binding sites target specific biological processes

To determine potential functions for the newly derived binding sites, we tested whether genes predicted to be targeted by binding sites with recent origins in hominids were involved in specific biological processes or pathways. Such enrichment was determined for genes near hominid-specific binding sites compared to the total list of protein-bound sites for each factor, where each TFBS was mapped to the nearest TSS, up to a distance of 100kb. This allowed us to assess potential lineage-specific functions of these sites relative to sites of more ancient origin.

Genes located nearest to hominid-specific binding sites were more frequently enriched for neural and sensory-related functions, and were in many cases involved in neurological pathways (Table 3.3). CTCF, MYC, and SOX2 target gene sets were all enriched for GO categories involved in sensory perception, while GATA1, MYC, ETS1, and MAX were enriched for neural development and differentiation categories. Among the six factors, neural-related functions are only well-documented for SOX2, which is involved in neuronal-cell maintenance (Giorgetti et al., 2012; Cavallaro et al., 2008) and whose hominid-specific target sites are enriched genes involved in sensory perception. Similarly, genes in proximity to hominid-specific binding sites

for CTCF and MYC were enriched for sensory perception processes and pathways, particularly those related to olfaction, and in the case for MYC, hominid-specific target genes were also enriched for genes involved in synapse assembly and receptor clustering and binding. Hominid-specific binding sites for GATA1, most commonly known for its role in erythroid differentiation (Pevny et al., 1995), were also found enriched near genes involved in axon extension of neural cells. For ETS1, hominid-specific binding sites were near genes involved in spinal cord neuron differentiation, ventral spinal cord development, and behavioral fear response.

3.5 A TFBS turnover event

Using our framework, we then utilized genome-wide chromatin data to search for potential functional consequences driven by birth or death of specific lineage-specific TFBS. We intersected the lineage-specific TFBSs with predicted human enhancer regions marked by ChromHMM model (Ernst and Kellis, 2012) as well as *in vivo* verified enhancers listed in the VISTA Enhancer Browser (Visel et al., 2007). Figure 3.7 shows a potential functional take-over through TFBS turnover inside an enhancer after human-mouse divergence. At the sequence level, two MAX binding sites were identified by our method with an ancestral one and a primate-specific binding site emerging after human-bushbaby split (Figure 3.7). Here these two MAX binding sites are also MYC binding sites since MAX and MYC have very similar motif (their ChIP-seq peaks overlap in Figure 3.7). The orthologous region of predicted primate-specific MAX/MYC binding site has no MAX or MYC ChIP-seq signal at all in mouse, which is consistent with our lineage-specific prediction. Since the young MAX/MYC binding site only locates 1,700bp upstream of the ancestral one and ChIP-seq intensity of ancestral binding site is much weaker in human compared to mouse, this is likely to be a turnover of MAX/MYC binding site within the enhancer. Then we asked whether the function of predicted enhancer was conserved between human and mouse. Interestingly, despite the potential turnover of MAX/MYC binding site in the sequence level, the mouse orthologous region of predicted enhancer was found to drive reproducible LacZ expression in E11.5 mouse blood cell as demonstrated by *in vivo* transgenic mouse embryos essay based on VISTA

Enhancer Browser, which confirms that the predicted enhancer also functions as an enhancer in mouse. It certainly remains to be solved whether this enhancer regulates the same genes in human and mouse and why the ChIP-seq signal on ancient MAX/MYC binding site is much weaker than the younger one in human. Nevertheless, this example demonstrates the ability of our method to compare functional level dynamics with sequence level difference in an evolutionary framework.

3.6 Figures and Tables

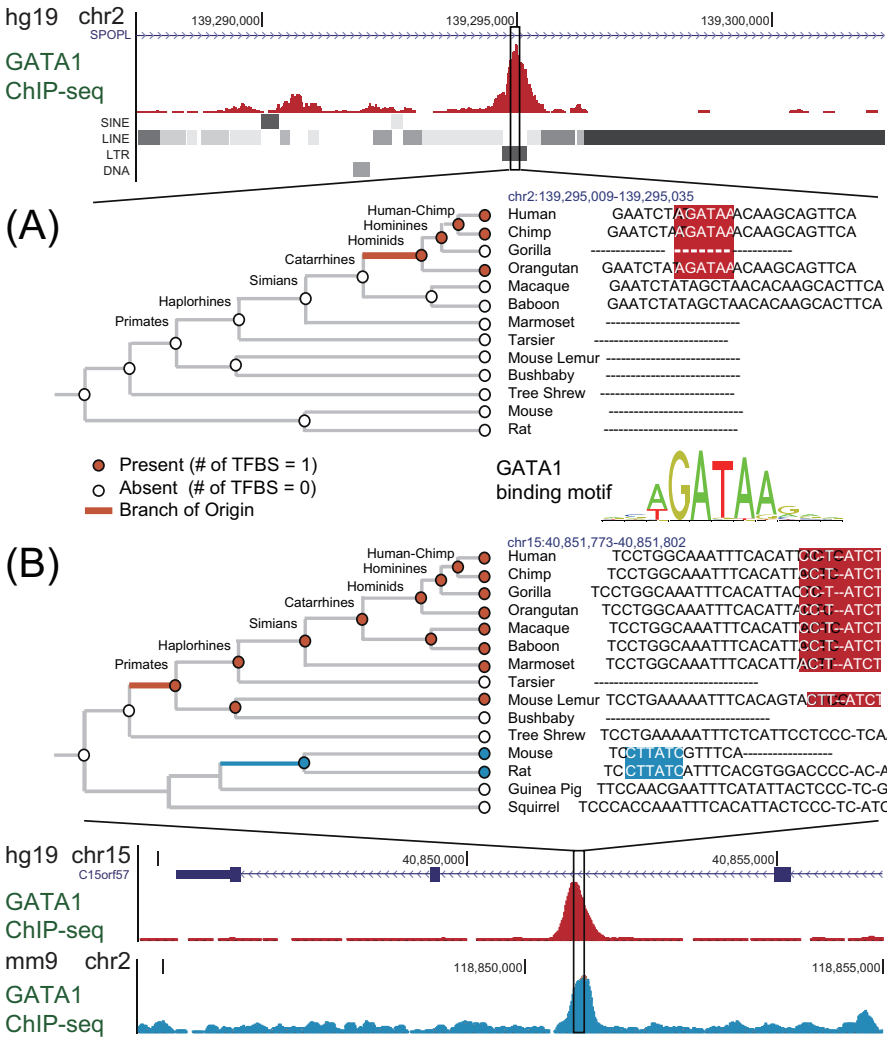

Figure 3.1: Model framework for lineage-specific GATA1 binding sites. Multiple alignments are shown for two GATA1-bound regions in humans. Red and blue boxes in the alignment correspond to GATA1 binding sites. Phylogenies illustrate the birth-death model framework, where the most likely number of binding sites is assigned to each ancestral node (denoted here as either present or absent, at values 1 or 0 in this example). Highlighted branches denote the branch of origin. Evolutionary comparisons were conducted across ten primate species, as well as 36 non-primate vertebrates (not all are shown). (A) A binding site originating within an LTR insertion. (B) A genomic region containing a human GATA1 binding site originating along the ancestral primate lineage and a GATA1 binding site specific to mouse and rat. Despite nearly identical locations of the ChIP-seq peaks across human and mouse (in analogous Erythroblast cell lines), the ability of the method to identify specific branches of origin allows us to identify cases of TFBS turnover in close proximity.

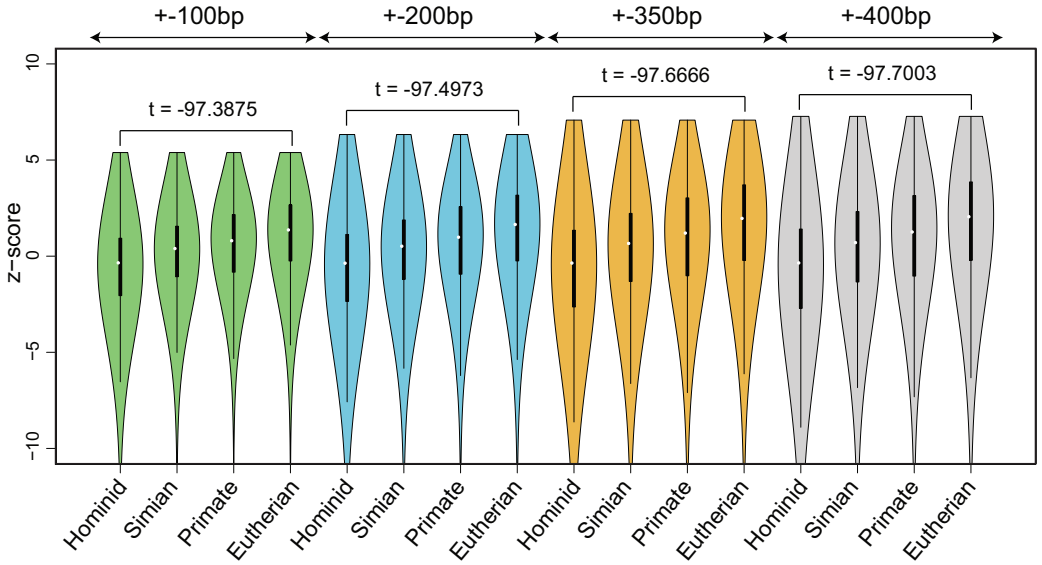

Figure 3.2: PhyloP conservation vs. TFBS with different branch of origins. We used the PhyloP mammalian conservation scores available at the UCSC Genome Browser to determine the sequence conservation level for TFBS with different branch of origins in human. X-axis shows TFBS with different branch of origins for four different window sizes surrounding the peak summit. Y-axis shows the Z-score distribution for each group. For a specific TF, we first computed the average PhyloP score (X) in each ChIP-seq peak and then calculated the average score (M) as well as standard deviation (SD) across all peaks in the genome. We then grouped the binding sites according to their branch of origin (in four groups: Hominid, Simian, Primate, and Eutherian). Finally, we calculated the Z-score, i.e. $(X - M)/SD$, in each age group. t-statistic from t-test between the youngest and the oldest for each group is also shown.

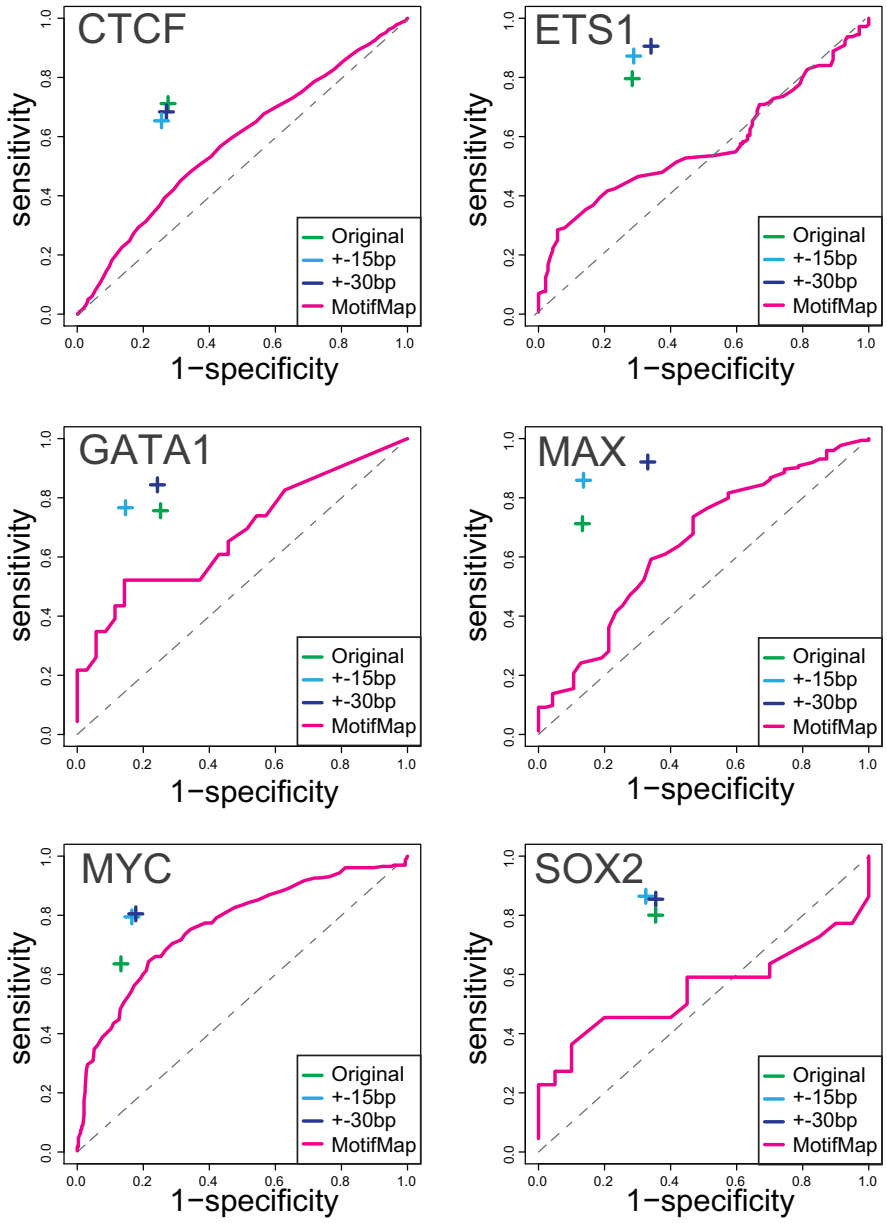

Figure 3.3: Comparison between our method and MotifMap. A receiver operating characteristic (ROC) curve shows the prediction power between our method and MotifMap. ROC curves for MotifMap were generated using different BBLS thresholds (ranging from zero to the maximum possible BBLS score here, 4.73) to call a TFBS as a conserved one. In our method, we tested two shift sizes, ±15 bp (light blue) and ±30 bp (dark blue). The results from MotifMap were based on ±15 bp shift size (magenta). See Section 2.4 for detailed explanation of the comparison method and how the benchmark dataset was constructed.

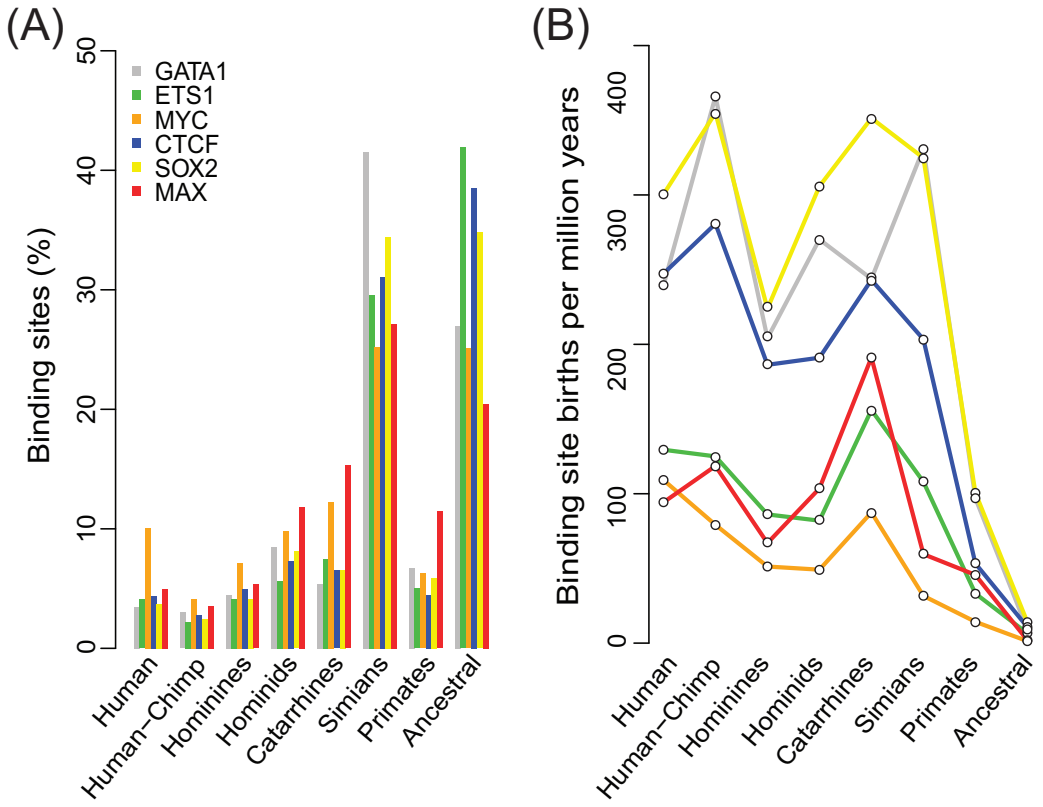

Figure 3.4: Time of origins for binding sites of six TFs in humans. Binding motifs were determined using human ChIP-seq data for GATA1, SOX2, MYC, CTCF, ETS1, and MAX. The branch of origin was determined for each binding site within the (−100,+100) region relative to a human ChIP-seq peak summit. (A) Distribution of the branch of origin for each binding site. Branch labels correspond to those in Figure 3.1. Ancestral binding sites have origins prior to human-mouse divergence. (B) The rate of binding site creation along branches ancestral to humans. Rates were estimated by dividing the number of sites originating along each branch by evolutionary time, including only binding sites currently existing in humans.

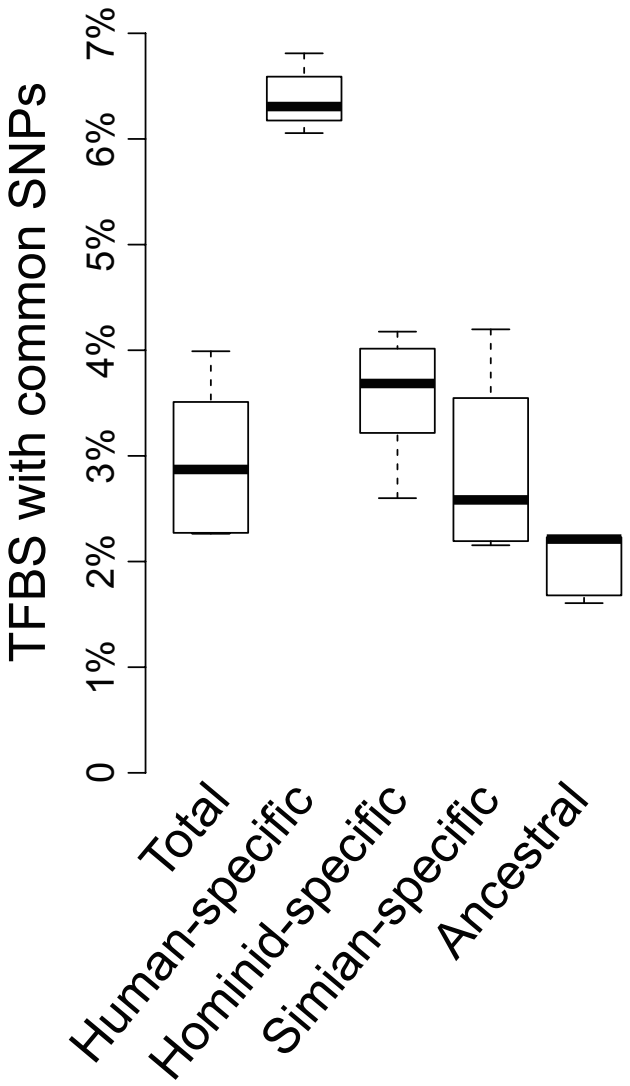

Figure 3.5: Within-species variation of binding sites according to time of origin. Boxplots show the fraction of TFBSs containing common SNPs of human population (Sherry et al., 2001), where plots show the median (center line), upper- and lower-quartile (boxes), and range (whisker extremes) of percentages across the TFBSs of six TFs. TFBSs are categorized as human-specific, hominid-specific (not including human-specific sites), Simian primate-specific (not including hominid-specific sites), and ancestral (present in the human-mouse common ancestor). Overall fractions (including all sites) are shown in the left-most boxplot. Note the substantial rise in the amount of human variation within more recently derived binding sites compared to older sites.

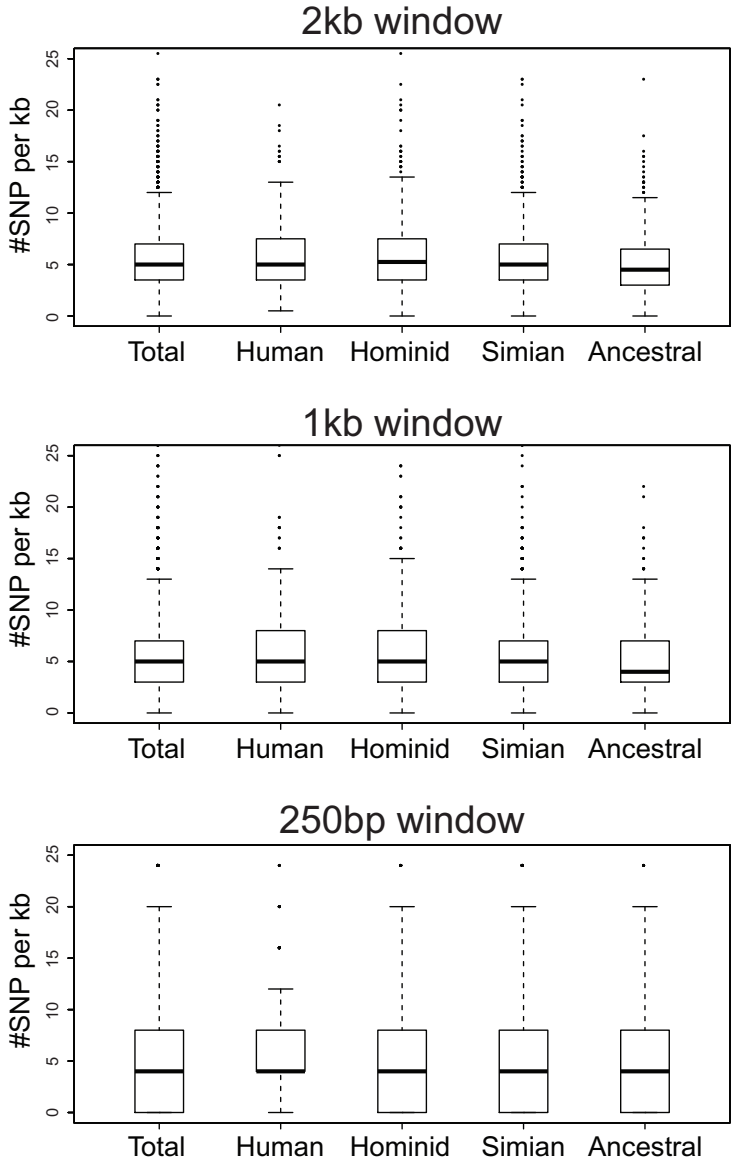

Figure 3.6: Background SNP density for TFBSs with different branch of origin. TFBSs were grouped into different branches of origin (X-axis). To calculate the background SNP density surrounding these TFBSs, we extended 1 kb, 500 bp, or 125 bp to both directions (i.e., 2k, 1k, or 250 bp window) and counted the number of common SNPs in flanking windows. The figure shows that there are no significant differences of SNP density surrounding the TFBSs with different branches of origin.

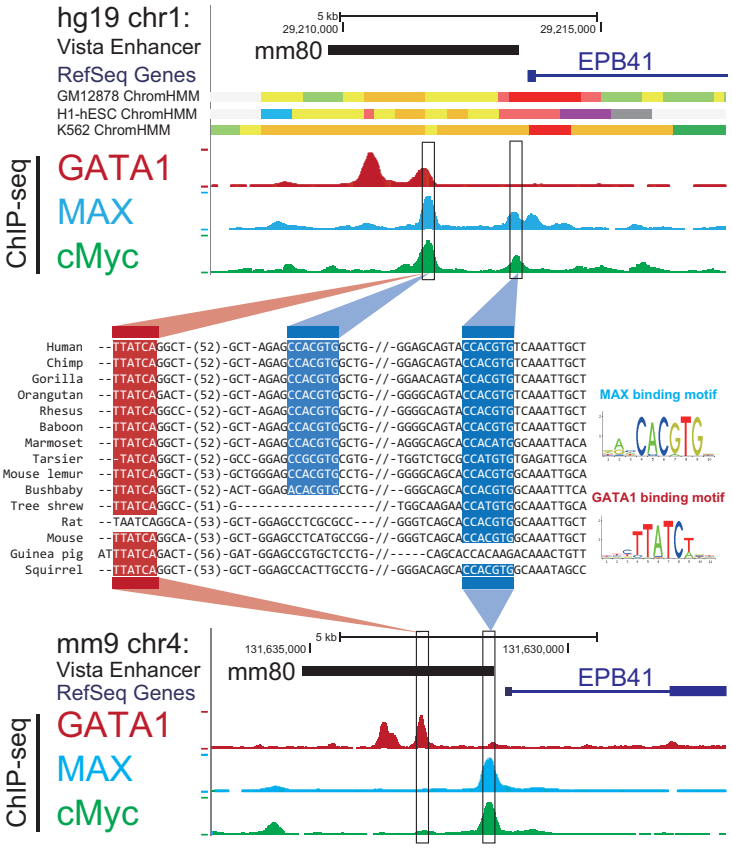

Figure 3.7: A TFBS turnover event within a functionally conserved enhancer. A TFBS turnover event shows the impact of lineage-specific TFBS within an enhancer. The Genome Browser view shows the upstream of human gene EPB41. VISTA Enhancer track and ChromHMM track (orange means strong enhancer, yellow means weak enhancer) indicate a putative human enhancer. ChIP-seq signals of three TFs used in this study near predicted enhancer region are consistent with predicted lineage-specific binding site represented by 46-way multiple sequence alignment (only a subset of species are shown). Note that here the two MAX binding sites are also MYC binding sites since MAX and MYC have very similar motif. A potential TFBS turnover is observed between two predicted MAX/MYC binding sites (1700 bp apart). Different TFBSs are highlighted in different colors with MAX in blue and GATA1 in red. The predicted enhancer may function as blood cell specific enhancer in mouse, demonstrated by images of LacZ positive E11.5 mouse transgenic embryo on the VISTA Enhancer Browser (Visel et al., 2007) (ID: mm80;`http://enhancer.lbl.gov/cgi-bin/imagedb3.pl?form=presentation&show=1&experiment\_id=80\&organism\_id=2`).

Table 3.1: Performance comparison with MotifMap and PReMod. We compared our method with phylogenetic footprinting methods MotifMap for element level comparison and PReMod for module level comparison. For MotifMap, BBLS score cutoff to call a conserved TFBS is 1.91 in this table. Bold numbers indicate the best among the three methods. PReMod database does not contain results for CTCF and SOX2, so we filled 'NA'. Here we tested our method using two window sizes: ±15bp and ±30bp. The window size used by MotifMap is ±15bp. The overall accuracy (which balances sensitivity and specificity) of our method is much better than MotifMap and PReMod. We also show the comparison between our method and MotifMap using different BBLS thresholds and for all TFs our method outperformed MotifMap (Figure3.3).

(a) Results from ± 15bp shift size in our method

	Sensitivity			Specificity			Accuracy		
	Our method	MotifMap	PReMod	Our method	MotifMap	PReMod	Our method	MotifMap	PReMod
CTCF	**0.6533**	0.3115	NA	0.7405	0.7940	NA	**0.7100**	0.512	NA
ETS1	**0.8722**	0.5903	0.3633	**0.7074**	0.3669	0.7005	**0.7381**	0.4806	0.4989
GATA1	**0.7664**	0.4348	0.4054	0.8496	**0.8857**	0.6974	**0.8464**	0.7069	0.5959
MAX	**0.8591**	0.4368	0.5202	**0.8596**	0.7447	0.6957	**0.8595**	0.5023	0.5758
MYC	0.7944	**0.8478**	0.3448	**0.8304**	0.4658	0.7802	**0.8215**	0.6341	0.5362
SOX2	**0.8643**	0.0909	NA	0.6708	**1.0000**	NA	**0.6783**	0.5238	NA

(b) Results from ± 30bp shift size in our method

	Sensitivity			Specificity			Accuracy		
	Our method	MotifMap	PReMod	Our method	MotifMap	PReMod	Our method	MotifMap	PReMod
CTCF	**0.6836**	0.3115	NA	0.7257	0.7940	NA	**0.7110**	0.5120	NA
ETS1	**0.9055**	0.5903	0.3633	**0.6550**	0.3669	0.7005	**0.7017**	0.4806	0.4989
GATA1	**0.8437**	0.4348	0.4054	0.7529	**0.8857**	0.6974	**0.7565**	0.7069	0.5959
MAX	**0.9208**	0.4368	0.5202	0.6649	**0.7447**	0.6957	**0.6934**	0.5023	0.5758
MYC	0.8049	**0.8478**	0.3448	**0.8184**	0.4658	0.7802	**0.8151**	0.6341	0.5362
SOX2	**0.8543**	0.0909	NA	0.6407	**1.0000**	NA	**0.6490**	0.5238	NA

Table 3.2: Evaluation with human-mouse ChIP-seq factor bound regions

		Shared peaks (human-mouse)[b]		Lineage-specific peaks[c]	
Factor/Motif	Category[a]	Ancestral Sites[d]	Lineage-specific Sites[e]	Ancestral Sites[d]	Lineage-specific Sites[e]
GATA1	**Human (Total)**	**433 (73.3%)**	**158 (26.7%)**	**6921 (34.9%)**	**12914 (65.1%)**
AGATAAG	*With Mouse*	*88 (14.9%)*	*24 (4.10%)*	*985 (5.00%)*	*623 (3.10%)*
	TFBS-pair aligned	*34 (5.80%)*	*11 (1.90%)*	*358 (1.80%)*	*93 (0.50%)*
SOX2	**Human (Total)**	**340 (75.7%)**	**109 (24.3%)**	**9451 (47.4%)**	**10487 (52.6%)**
WTAACAA	*With Mouse*	*123 (27.4%)*	*26 (5.80%)*	*2242 (11.2%)*	*1009 (5.10%)*
	TFBS-pair aligned	*85 (18.9%)*	*7 (1.60%)*	*747 (3.80%)*	*138 (0.70%)*
MYC	**Human (Total)**	**406 (58.5%)**	**288 (41.5%)**	**704 (26.4%)**	**1966(73.6%)**
KCACGTG	*With Mouse*	*111 (16.0%)*	*39 (5.60%)*	*109 (4.10%)*	*93 (3.50%)*
	TFBS-pair aligned	*64 (9.20%)*	*16 (2.30%)*	*55 (2.10%)*	*38 (1.40%)*
MAX	**Human (Total)**	**420 (69.8%)**	**182 (30.2%)**	**1167 (21.2%)**	**4350 (78.9%)**
KCACGTG	*With Mouse*	*134 (22.3%)*	*49 (8.10%)*	*209 (3.80%)*	*269 (4.90%)*
	TFBS-pair aligned	*79 (13.1%)*	*19 (3.20%)*	*125 (2.30%)*	*66 (1.20%)*
ETS1	**Human (Total)**	**644 (73.5%)**	**232 (26.5%)**	**3885 (48.1%)**	**4189 (51.9%)**
MGGAAGT	*With Mouse*	*172 (19.6%)*	*35 (4.00%)*	*705 (8.70%)*	*283 (3.50%)*
	TFBS-pair aligned	*81 (9.3%)*	*9 (1.00%)*	*283 (3.50%)*	*58 (0.70%)*
CTCF	**Human (Total)**	**2008 (68.0%)**	**947 (32.0%)**	**3829 (41.2%)**	**5458 (58.8%)**
GGGGCKC	*With Mouse*	*766 (25.9%)*	*277 (9.40%)*	*770 (8.30%)*	*323 (3.50%)*
	TFBS-pair aligned	*256 (8.70%)*	*73 (2.50%)*	*231 (2.50%)*	*42 (0.50%)*

[a]The first row in each section gives the total number of ChIP-seq peaks with binding sites in humans within the (100,+100) window separated into categories. The second row shows the number of these peaks also containing a TFBS in the orthologous regions in mouse, while the third row gives the number of aligned binding sites across the two species. Percentages are given with respect to the total number of shared and lineage-specific ChIP-seq peaks for each factor.
[b]Shared peaks are human ChIP-seq peaks within 200 bp of a ChIP-seq peak summit in the orthologous region in mouse. Analogous cell types were used across species (GATA1: Erythroblasts, SOX2: Embryonic stem cells, MYC, MAX, ETS1, CTCF: B-lymphocytes).
[c]Lineage-specific peaks in human are not within 200 bp of a mouse ChIP-seq peak in the analogous cell type. Only human peaks with identifiable orthologous regions in mouse were included.
[d]The numbers (and fractions) of human ChIP-seq peaks in each category containing binding motif occurrences estimated to be present in the human-mouse ancestor (Ancestral sites).
[e]The numbers (and fractions) of human ChIP-seq peaks in each category containing only binding motif occurrences originating after human-mouse divergence (Lineage-specific sites).

Table 3.3: Gene functions and pathways associated with hominid-specific TFBS. Shown are the top-ranking biological processes and gene pathways for genes associated with hominid-specific binding sites for each TF. Functional category enrichment was determined relative to the target genes for the comprehensive list of binding sites, with P-values and fold-enrichment over this background set of target genes determined by GREAT.

Factor	Biological Process	P-val	Fold	Biological pathways	P-val
CTCF	Positive regulation of actin filament polymerization (32 genes)	3e-14	3.45x	Olfactory signaling pathway (346 genes)	7e-5
	Retrograde transport endosome to Golgi (29 genes)	2e-12	3.50x	Phase II conjugation (37 genes)	4e-3
	Positive regulation of protein polymerization (44 genes)	2e-8	2.41x	Stearate biosynthesis I (animals) (12 genes)	1e-2
	Detection of chemical stimulus involved in sensory perception of smell (385,genes)	5e-8	2.20x	Regulation of lipid metabolism by peroxisome proliferator-activated receptor alpha (PPARalpha) (32 genes)	1e-2
GATA1	Inositol phosphate metabolic process (16 genes)	2e-4	2.40x	Ketone body metabolism (5 genes)	6e-4
	Positive regulation of histone acetylation (12 genes)	6e-4	2.90x	Mevalonate pathway I (10 genes)	6e-4
	Axon extension involved in axon guidance (14 genes)	1e-3	5.23x	Transmission across electrical synapses (5 genes)	2e-3
	Organophosphate catabolic process (11 genes)	2e-3	2.77x	Tryptophan degradation III (eukaryotic) (9 genes)	5e-3
MYC	Synapse assembly (41 genes)	2e-5	3.33x	Olfactor signaling pathway (346 genes)	4e-5
	Sensory perception of chemical stimulus (456 genes)	7e-5	2.11x	Neurotransmitter receptor binding and downstream transmission in the postsynaptic cell (90 genes)	6e-4
	Receptor clustering (17 genes)	7e-5	3.33x	NCAM1 interactions (23 genes)	2e-3
	Neuron maturation (25 genes)	1e-4	3.30x	CREB phosphorylation through the activation of Ras (29 genes)	2e-3
SOX2	rRNA processing (109 genes)	2e-4	2.01x	Signaling by Aurora kinases (98 genes)	2e-4
	Detection of stimulus involved in sensory perception	5e-4	1.91x	Aurora B signaling (41 genes)	2e-4
	GTP metabolic process (249 genes)	8e-4	1.50x	The citric acid (TCA) cycle and respiratory electron transport (106 genes)	2e-4
	tRNA modification (22 genes)	1e-3	3.81x	Eukaryotic translation elongation (88 genes)	5e-4
ETS1	Ventral spinal cord development (26 genes)	2e-4	3.55x	Mitotic spindle checkpoint (19 genes)	4e-4
	Receptor guanylyl cyclase signaling pathway (11 genes)	2e-3	4.51x	APC-Cdc20 mediated degradation of Nek2A (23 genes)	8e-4
	Cell differentiation in spinal cord (37 genes)	3e-3	2.70x	Phosphorylation of Emi1 (20 genes)	6e-3
	Behavioral fear response (15 genes)	4e-3	2.80x	Tetrahydrobiopterin (BH4) synthesis, recycling, salvage and regulation (12 genes)	1e-2
MAX	Establishment of organelle localization (100 genes)	1e-3	3.36x	Signal amplification (16 genes)	5e-3
	Neural crest cell differentiation (57 genes)	3e-3	2.14x	Thrombin signaling through proteinase activated receptors (PARs) (17 genes)	5e-3
	Neural crest cell development (50 genes)	7e-3	2.06x	PAR4-mediated thrombin signaling events (15 genes)	1e-2
	Positive regulation of lipid transport (19 genes)	8e-3	2.20x	Signaling by Robo receptor (23 genes)	1e-2

CHAPTER 4

CONCLUSION

Understanding the evolution of the cis-regulatory elements is an essential step toward understand the phenotype difference across species, specially for close related species (Wittkopp and Kalay, 2012). Studies regarding the evolution of TFBS can largely be separated into those emphasize cross-species conservation of cis-regulatory elements and those highlighting the substantial divergence of TFBS. To some extent, this dichotomy may largely reflect the difference between experimental methods and *in silico* methods. Despite some studies have inferred the lineage-specific evolution of TFBS (Lindblad-Toh et al., 2011; Lowe et al., 2011; Hiller et al., 2012), all of them are based on base-by-base details of MSA, which may lead to unreliable predictions because of the low-quality issue of MSA in non-coding region (McLean et al., 2011; Chen and Tompa, 2010; Kim and Ma, 2011). There are pressing computational challenge to understand the history of cis-regulatory element such as TFBS. In addition, it has long argued that alterations in non-coding regions are responsible for many, if not most, species-specific traits (Wray, 2007; Davidson, 2001; King and Wilson, 1975). Thus, in this work we presented an initial step using *in silico* methods to model the evolution of cis-regulatory elements without relying on accurate cross-species alignment.

Applying our method to six human TF ChIP-seq revealed that a high fraction of TFBSs have origins after human-mouse split (Figure 3.4), which is consistent with previous *in vivo* cross-species comparison in a limited number of species (Odom et al., 2007; Schmidt et al., 2010) and a human-mouse comparison we did here (Table 3.2). This observation may not be a specific phenomena because of the six TFs we picked as motifs of those six are among the most conserved TF motif between human and mouse. The fact that a motif is conserved does not mean majority of binding sites are conserved according to our finding. Next, we compared genomic functional patterns across TFBS with different branch of origin. Younger TFBS have higher

density of common SNP (Figure 3.5) and their target genes are more enriched in neural related functions and pathways (Table 3.3). This results can help understand the roles of lineage-specific TFBS in shaping gene regulation across different species.

A natural future direction for this work would be as follows. First, we can expand our analysis to other TFs, since there are an estimated 1700 1900 TFs in the human genome (Vaquerizas et al., 2009) and dozens of them have ChIP-seq data available in ENCODE project (ENCODE Project Consortium, 2011). Second, it would be interesting if we can determine the specific regulatory effects of the recently derived TFBSs identified using this method. For instance, enrichment for within-species variation among recently derived binding sites raises the intriguing possibility that recently derived TFBSs most responsible for phenotypic differences across species are also the elements responsible for within-species variation. Future work will be necessary to demonstrate whether this is the case and is this pattern still hold to somatic mutations in cancer patients. Also, our current model needs to be integrated with gene expression data to understand the interplay between cis-regulatory element evolution (e.g., binding site turnover and lineage-specific sites) and gene expression differences across different species (Tirosh and Barkai, 2011; Tirosh et al., 2008; Romero et al., 2012; Cusanovich et al., 2014). Next, the whether lineage-specific TFBSs are more tissue-specific should also to inspected. Last, how to integrate TFBS within enhancer or promoter together and study the evolution of those large cis-regulatory modules is also an interesting direction.

Overall, we believed our birth-death probabilistic model would be highly useful to comprehensively under the evolution of genome-wide TFBS. By add a time dimension onto the current human genome annotation resources we could understand human genome better.

REFERENCES

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Blanchette, M., Bataille, A. R., Chen, X., Poitras, C., Laganire, J., Lefbvre, C., Deblois, G., Gigure, V., Ferretti, V., Bergeron, D., Coulombe, B. and Robert, F. (2006). Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Research *16*, 656–668.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D. and Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research *14*, 708–715.

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M. and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. Science (New York, N.Y.) *317*, 815–819.

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H. and Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Research *18*, 1752–1762.

Cavallaro, M., Mariani, J., Lancini, C., Latorre, E., Caccia, R., Gullo, F., Valotta, M., DeBiasi, S., Spinardi, L., Ronchi, A., Wanke, E., Brunelli, S., Favaro, R., Ottolenghi, S. and Nicolis, S. K. (2008). Impaired generation of mature neurons by neural stem cells from hypomorphic Sox2 mutants. Development (Cambridge, England) *135*, 541–557.

Cavender, J. A. (1978). Quasi-Stationary Distributions of Birth-and-Death Processes. Advances in Applied Probability *10*, 570–586.

Cavener, D. R. (1987). Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. Nucleic Acids Research *15*, 1353–1361.

Chen, X. and Tompa, M. (2010). Comparative assessment of methods for aligning multiple genome sequences. Nature Biotechnology *28*, 567–572.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L. and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. PLoS genetics *10*, e1004226.

Davidson, E. H. (2001). Genomic Regulatory Systems: In Development and Evolution. 1st edition edition, Academic Press, San Diego.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B *39*, 1–38.

ENCODE Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology *9*, e1001046.

Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature Methods *9*, 215–216.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics *25*, 471–492.

Giorgetti, A., Marchetto, M. C. N., Li, M., Yu, D., Fazzina, R., Mu, Y., Adamo, A., Paramonov, I., Cardoso, J. C., Monasterio, M. B., Bardy, C., Cassiani-Ingoni, R., Liu, G.-H., Gage, F. H. and Izpisua Belmonte, J. C. (2012). Cord blood-derived neuronal cells by ectopic expression of Sox2 and c-Myc. Proceedings of the National Academy of Sciences of the United States of America *109*, 12556–12561.

Hardison, R. C., Oeltjen, J. and Miller, W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Research *7*, 959–966.

He, X., Ling, X. and Sinha, S. (2009). Alignment and Prediction of cis - Regulatory Modules Based on a Probabilistic Model of Evolution. PLOS Comput Biol *5*, e1000299.

Hiller, M., Schaar, B. T., Indjeian, V. B., Kingsley, D. M., Hagey, L. R. and Bejerano, G. (2012). A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. Cell Reports *2*, 817–823.

Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, N.Y.) *316*, 1497–1502.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D. and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Research *32*, D493–496.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M.-Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O. and Snyder, M. (2010). Variation in transcription factor binding among humans. Science (New York, N.Y.) *328*, 232–235.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254.

Kheradpour, P., Stark, A., Roy, S. and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Research *17*, 1919–1931.

Kim, J. and Ma, J. (2011). PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucleic Acids Research *39*, 6359–6368.

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. Science (New York, N.Y.) *188*, 107–116.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alfldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin, J., Bloom, T., Chin, C. W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. A., Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree, A., Dihn, H. H., Fowler, G., Jhangiani,

S., Joshi, V., Lee, S., Lewis, L. R., Nazareth, L. V., Okwuonu, G., Santibanez, J., Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Genome Institute at Washington University, Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., Sodergren, E., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard, K. S., Pedersen, J. S., Lander, E. S. and Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B. and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., Kingsley, D. M., Lindblad-Toh, K. and Haussler, D. (2011). Three periods of regulatory innovation during vertebrate evolution. Science (New York, N.Y.) *333*, 1019–1024.

Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A. and Hein, J. (2007). Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. Genome Research *18*, 000–000.

Majoros, W. H. and Ohler, U. (2010). Modeling the Evolution of Regulatory Elements by Simultaneous Detection and Alignment with Phylogenetic Pair HMMs. PLOS Comput Biol *6*, e1001037.

Margulies, E. H., Blanchette, M., NISC Comparative Sequencing Program, Haussler, D. and Green, E. D. (2003). Identification and characterization of multi-species conserved sequences. Genome Research *13*, 2507–2518.

Matys, V., Fricke, E., Geffers, R., Gssling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Mnch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Research *31*, 374–378.

McLean, C. Y., Reno, P. L., Pollen, A. A., Bassan, A. I., Capellini, T. D., Guenther, C., Indjeian, V. B., Lim, X., Menke, D. B., Schaar, B. T., Wenger, A. M., Bejerano, G. and Kingsley, D. M. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature *471*, 216–219.

Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S. L., Nekrutenko, A., Giardine, B., Harris, R. S., Tyekucheva, S., Diekhans, M., Pringle, T. H., Murphy, W. J., Lesk, A., Weinstock, G. M., Lindblad-Toh, K., Gibbs, R. A., Lander, E. S., Siepel, A., Haussler, D. and Kent, W. J. (2007). 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Research *17*, 1797–1808.

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K. and Fraenkel, E. (2007). Tissue-Specific Transcriptional Regulation has Diverged Significantly between Human and Mouse. Nature genetics *39*, 730–732.

Pevny, L., Lin, C. S., D'Agati, V., Simon, M. C., Orkin, S. H. and Costantini, F. (1995). Development of hematopoietic cells lacking transcription factor GATA-1. Development (Cambridge, England) *121*, 163–172.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research *20*, 110–121.

Romero, I. G., Ruvinsky, I. and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics *13*, 505–516.

Sandelin, A., Alkema, W., Engstrm, P., Wasserman, W. W. and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research *32*, D91–94.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E. V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C. and Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. Nature *483*, 169–175.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P. and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science (New York, N.Y.) *328*, 1036–1040.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Research *29*, 308–311.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W. and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research *15*, 1034–1050.

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Harvard FlyBase curators, Berkeley Drosophila Genome Project, Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-W., Han, M. V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M. and Kellis, M. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature *450*, 219–232.

Tirosh, I. and Barkai, N. (2011). Inferring regulatory mechanisms from patterns of evolutionary divergence. Molecular Systems Biology *7*, 530.

Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M. and Barkai, N. (2008). On the relation between promoter divergence and gene expression evolution. Molecular Systems Biology *4*, 159.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. Nature Reviews. Genetics *10*, 252–263.

Visel, A., Minovitsky, S., Dubchak, I. and Pennacchio, L. A. (2007). VISTA Enhancer Browser–a database of tissue-specific human enhancers. Nucleic Acids Research *35*, D88–92.

Ward, L. D. and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science (New York, N.Y.) *337*, 1675–1678.

Wittkopp, P. J. and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics *13*, 59–69.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics *8*, 206–216.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. and Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. Molecular Biology and Evolution *20*, 1377–1419.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S. and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338–345.

Xie, X., Mikkelsen, T. S., Gnirke, A., Lindblad-Toh, K., Kellis, M. and Lander, E. S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proceedings of the National Academy of Sciences of the United States of America *104*, 7145–7150.

Xie, X., Rigor, P. and Baldi, P. (2009). MotifMap: a human genome-wide map of candidate regulatory motif sites. Bioinformatics (Oxford, England) *25*, 167–174.

Yokoyama, K. D. and Pollock, D. D. (2012). SP transcription factor paralogs and DNA-binding sites coevolve and adaptively converge in mammals and birds. Genome Biology and Evolution *4*, 1102–1117.