Title
Predicting developmental limb enhancers and quantifying motif sequence differences between enhancers.

Author
Friedrich, Tara

Publication Date
2017

Peer reviewed|Thesis/dissertation

Predicting developmental limb enhancers and quantifying motif sequence differences between enhancers.

by

Tara Friedrich

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UN FRANCISCO

Acknowledgements

Lots of people have helped me get to where I am today and I apologize if I have left anyone out. First, I would like to thank my mother and father for always pushing me to try harder and encouraging me to persist in my scientific career. Both of my parents encouraged me to think like a scientist from a young age. I don't think I would be here today without that early upbringing.

Joining the Pollard lab was the best decision I made in graduate school. I can't emphasize the importance of finding a good mentor in school. My advisor, Katherine Pollard, provided me with a model of leadership that I will carry with me for the rest of my career. Additionally, I would like to emphasize that the lab is full of driven individuals that have supported me in all my scientific endeavors. Fellow graduate students, Aram Avila-Herrera and Genevieve Erwin Haliburton, guided much of my direction during the early years. In addition, postdoctoral scholars Nandita Garud, Hassan Samee, Patrick Bradley, and Geoffrey Fudenberg were key in helping me prepare for my future steps in my career.

I want to thank my committee members for sitting through long meetings and giving me the feedback I needed. Nadav Ahituv, Benoit Bruneau, and Jeff Wall have given me valuable advice on whether to pursue certain directions in my research.

I also want to acknowledge all the individuals that supported me personally in school. In particular, I have always valued Sara Calhoun's well-thought-out advice and Rose

Citron's insight when making important decisions. I really enjoyed interacting with members of the Ahituv lab, particularly Hane Ryu. And I often felt like an extended member of the Srivastava lab, thanks to Nicole Stone's efforts to incorporate me. And where would I be without coffee breaks with Alex Williams, Svetlana Lyalina and Kathleen Keough? I really appreciate the moral support!

Predicting developmental limb enhancers and quantifying motif sequence differences between enhancers

by

Tara Friedrich

Abstract

Gene regulation can contribute to phenotypic divergence across species and cell types. By comparing regulatory regions between cell types and between species we can gain an understanding of how sequence changes affect gene regulation and ultimately organismal phenotypes and disease. Using computational methods, I quantified motif enrichment between sets of enhancers in order to characterize functional differences. I was able to identify transcription factors that showed a significant difference in the number of motifs enriched in homologous mouse and human cardiomyocyte enhancers. I also identified differentially enriched transcription factor motifs in embryonic stem cells and differentiated cardiomyocytes. These same methods were also applied to a third dataset in order to detect differences between binding sites that were unique to mutant SOX2 and binding sites that were shared between wildtype and mutant SOX2 binding sites. I found significant depletion of the OCT4:SOX2 motif in mutant SOX2 binding sites. In addition to this, my work also used a comparative genomics approach to identify regions that evolved rapidly in the bat ancestor, but are highly conserved in other vertebrates. I discovered 166 bat accelerated regions (BARs) that overlap epigenetic marks in developing mouse limbs and validated their function in limb development. Of particular note was an enhancer near the HoxD cluster that shows forelimb specific expression in bats compared to mice.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Big Picture: Genetic regulation of phenotype

Tissues in the human body are composed of cells that are regulated by DNA. From the early stages of development, regulatory regions found within DNA become active and exposed to transcription factors that can regulate genes. Transcription factors (TF) are proteins that bind DNA and activate or repress that gene. The order in which these transcription factors interact with their target DNA sequences over developmental time is essential for proper progression of development.

These regulatory pathways can be modified to produce slightly different phenotypes. Modifications can manifest in different ways. For example, genetic variation in non-genic regions can cause genes to be regulated differently across individuals. These expression changes can be ubiquitous or restricted to specific cell types, depending on the function of the mutated regulatory element. The Genotype-Tissue Expression (GTEx) Project was designed to study the relationship among genetic variation, gene expression, and other molecular phenotypes in multiple human tissues (Consortium et al., 2015). The researchers in this study observed how transcription varies among tissues as well as how truncated protein variants affect expression across tissues. They identified multiple expression quantitative trait loci (eQTLs) per gene, unique or shared among tissues in different individuals and positively correlated with the number of transcripts per gene. These differences can affect how different developmental processes turn on at different times thus resulting in differences in phenotypes.

Comparing genetic variation across species is a method of understanding how differences in phenotypes arose between related species. Although coding variants have been shown to cause species-specific phenotypes, it has been postulated and shown empirically that non-coding variation plays an equal or even greater role in divergence of sister taxa. One reason is that deleterious noncoding variants affecting the expression of a gene in a specific tissue would be more tolerated than deleterious mutations destroying the protein in all tissues. Because a large fraction of evolutionary innovation occurs in noncoding sequence, it is hypothesized that these non-coding variants allow for fine-tuning the regulation of specific genes under certain conditions without overall changing the function of the protein. We see examples of this in various vertebrates. For example, researchers were able to identify both coding and non-coding variants in Stickleback fish that are predictive of phenotypic differences between freshwater and marine species (Jones et al., 2012).

In addition, comparisons of the genomes of domesticated pigs and wild boars demonstrate multiple points about selection mechanisms and biological traits (Rubin et al., 2012). This study found an excess of derived nonsynonymous substitutions in domestic pigs. The authors suggest that these substitutions could be a result of positive selection and relaxed purifying selection after domestication. Three genes (*NR6A1*, *PLAG1*, and *LCORL*) at different loci together could identify the genetic source of vertebrae elongation in the domestic pig. *PLAG1* and *LCORL* also control stature in other domestic animals and in humans.

We also see how these noncoding variants can explain phenotypic differences between human populations. Researchers compared the genomes of indigenous peoples of highland Tibet to Han people inhabiting lowlands and found eight SNPs that diverge between these closely related populations (Beall et al., 2010). These SNPs are located next to a gene called EPAS that encodes for a transcription factor. This transcription factor regulates the production of red blood cells and could control the amount of oxygen in the blood. They go on to suggest that low hemoglobin content is advantageous to the Tibet population because high concentrations of hemoglobin are a symptom of chronic mountain sickness, thus showing that there could be selection for certain advantageous traits.

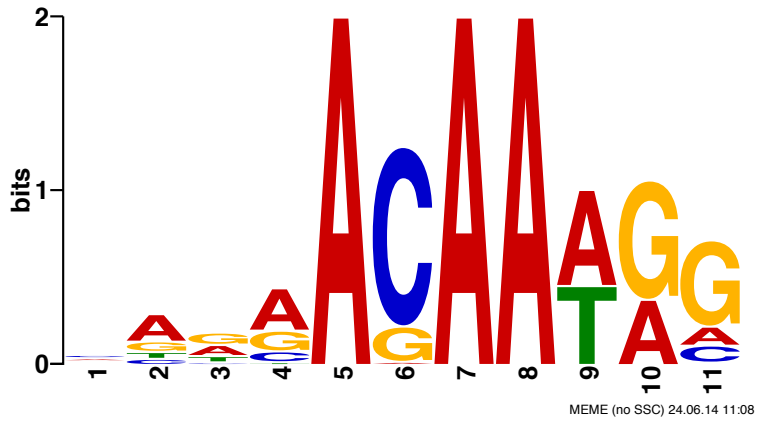
1.2 What is a transcription factor motif?

Regulatory genomic elements typically contain multiple motifs for one or more TFs. The TF proteins bind to these motif sequences to combinatorially modulate the expression of nearby genes (Maston, Landt, Snyder, & Green, 2012). TF motifs are to some extent degenerate (i.e., mutations away from the consensus sequence are tolerated), and therefore they are typically represented as probability distributions over nucleotides (A, C, G, and T) at each position in the motif (Stormo, 2000). For each TF, this distribution can be represented as position specific probability matrix (PSPM) that represents the occurrence of each nucleotide at each position (Figure 1.1). While TF binding depends on more than just the target DNA sequence (TF concentration, open chromatin, etc.), and even though the binding affinity of a TF towards a stretch of nucleotides is quantitative rather than binary, the presence or absence of TF motifs can be represented as a binary

event by scoring how well a sequence matches a TF's PSPM (details below). Because sequence changes can alter how well DNA matches a PSPM, mutations and substitutions can create or destroy motif instances.

Figure 1.1 Motif logo.

This motif logo displays the position-specific probability matrix (PSPM) representing the probability of each nucleotide occurrence at each position.



A typical approach to identify TF motifs in DNA sequences is to scan a sequence one position at a time using a PSPM and predict a motif at any position where the likelihood of a motif-length sub-sequence under the PSPM model is significantly higher than under a background distribution (Rahmann, Muller, & Vingron, 2003).

1.3 How can we detect regulatory regions?

Various pieces of information can be used to identify regulatory regions. For example, thousands of genomes have been sequenced in the past decade to draw conclusions about

sequence function (Alföldi & Lindblad-Toh, 2013). Conservation across species is a good indicator of functional importance. This is due to the fact that most change between species is the result of mutations with little functional impact. Consequently, mutations that fall in functional regions can be deleterious to the organism. Highly conserved non-coding regulatory elements are frequently gene regulatory elements (Harmston, Baresic, & Lenhard, 2013). However, the converse is not necessarily true: non-conserved sequence could, in fact, be functional and may point to species-specific phenotypes.

In addition to evolutionary conservation, regulatory marks are often used to identify regions that are active in cells. Researchers have used chromatin immunoprecipitation (ChIP) using antibodies directed at co-activators (CBP/p300) or at TFs, or histone marks that indicate presence of an regulatory regions (H3K27Ac, H3K4me1) (Sandmann et al., 2007; Bonn et al., 2012). Chromatin capture methods (3C, 4C, 5C, HiC) add complimentary information that identify distal regulatory elements that are in physical proximity to promoters of expressed genes (Shlyueva, Stampfel, & Stark, 2014). These techniques work by crosslinking and sequencing DNA with their targets in order to identify which regulatory regions interact which genes of interest. Techniques such as DNase-seq take advantage of the fact that regulatory regions are bound by TFs that block nucleosomes from binding (Crawford et al., 2006). Finally, algorithms can combine epigenetic information to define the boundaries of these regulatory regions (Hoffman et al., 2013). Studies have combined expression data with epigenetic information and found certain histone modification marks correlate with activation of gene expression ("An integrated encyclopedia of DNA elements in the human genome," 2012). These

approaches provide complimentary information for understanding how regulatory regions are defined and how they might functionally interact with their targets.

1.4 What is an enhancer?

Enhancers are a class of regulatory elements that are located distal to the genes they regulate. Enhancers regulate spatiotemporal gene expression and therefore play an important role in vertebrate development (Visel, Rubin, & Pennacchio, 2009). Because enhancers can function independently of the distance and orientation to their target, the enhancer sequence can be tested in a reporter assay *in vivo* which indicates whether the enhancer is active at that timepoint in development (Kvon, 2015). These methods can indirectly tell the researcher that an enhancer is active. However, a negative result does not mean that the enhancer does not function within a different context or developmental time point. Methods to parallelize these assays now allow researchers to test thousands of candidate enhancers at once and quantify their activity (Melnikov et al., 2012).

1.5 Enhancer evolution across species

When an otherwise conserved regulatory element is lost or mutated in one species, it is highly likely that its function changes. Nucleotide changes, copy number variation (CNVs), and chromosomal aberrations within enhancers have been shown to lead to phenotypic differences, such as limb malformations (VanderMeer & Ahituv, 2011). For example, regulatory regions in the 5'*Hoxd* locus have been implicated in digit specification during mammalian autopod development and loss of interactions with these

regions can result in limb phenotypes, similar to *Hoxd10-Hoxd13* gene deletions (Montavon et al., 2011).

1.6 Hypothesis

By comparing regulatory regions between cell types and between species we can gain an understanding of how sequence changes affect gene regulation and ultimately organismal phenotypes and disease. To do so, I applied methods to identify sequence differences and changes in TF binding sites in candidate regulatory regions and to quantify differences in their regulatory potential.

2 Motif enrichment differences between regulatory regions

Sequence divergence is usually measured in numbers of DNA substitutions or model-based estimates of rates of substitutions. These measures do not account for whether or not substitutions create or destroy TF motifs and are not well suited to quantify functional divergence (Ritter et al., 2010). It is challenging to predict the effect of a single motif loss or gain on the function of a regulatory region, because a loss may be compensated for by a nearby gain. However, a large cumulative change in the number of motifs across a regulatory region can alter expression of nearby genes, potentially resulting in differences in organismal traits, such as disease susceptibility (Bradley et al., 2010; Spivakov et al., 2012).

2.1 Method to detect binding site turnover

MotifDiverge is a method produced in my lab that quantifies how changes to DNA sequences affect their TF motif composition, which is a more meaningful measure of functional divergence for regulatory regions. This method is useful for understanding when non-coding mutations affect or do not affect the function of regulatory sequences. While the core of our approach is independent of the specifics regarding TF motif modeling, we also developed methodology to estimate the distribution of the difference in motif counts between sequences for any TF that has a motif model in the form of a PSPM. The sequences may be homologous or not, because our approach does not require (but can make use of) a sequence alignment.

Not many methods can quantify the divergence between DNA sequences based on differences in motif counts. The primary challenge is that in most biologically meaningful settings the sequences are related through evolution (i.e., they are homologous), and therefore motif instances are correlated. The motifDiverge method can detect if the difference in the number of motifs between two sequences is significantly different and can be used as a way of quantifying functional differences between two sequences. For homologous sequences, we can ask if the difference between two sequences is significant considering the fact that the sequences are phylogenetically related.

2.2 Applications to detect binding site turnover

My work leverages motifDiverge to compare transcription factor binding potential in several different contexts. In particular, I used motifDiverge to compare motifs in regulatory regions across species, cell types, and conditions. Examples include homologous regulatory regions in human versus mouse cardiomyocytes, sets of regulatory regions with activating marks in different cell types, and comparisons of mutant versus wildtype cell lines. These applications highlight the usefulness of comparing total counts of motifs between different sequences and accounting for sequence composition and length. In each application, my goal was to create a list of TFs whose ability to bind two regulatory sequences was predicted to be different due to motif losses or gains. These TFs and the diverged regulatory regions are candidates for discovering the genetic basis for differences in gene regulation and phenotypes across species, cell types, and conditions.

My first two applications of motifDiverge compare binding potential of regulatory sequences active during cardiac development. I analyzed a collection of gene regulatory elements identified via ChIP-seq for the active enhancer-marking histone modification histone 3 lysine 27 acetylation (H3K27ac) by Wamstad *et al*. (2012). This study identified genomic sequences marked by H3K27ac in mouse embryonic stem cells (ESCs) and at several subsequent developmental time points along the differentiation of ESCs into cardiomyocytes (CMs), which are beating heart cells. Tissue development is a useful system for illustrating our approach, because active regulatory elements and TFs that are important for regulating gene expression differ across cell types dynamically during development. Theses results were published in Kostka, Friedrich, Holloway, & Pollard (2014).

My third application of motifDiverge compares regulatory elements bound by a TF in ESCs in the presence and absence of a protein coding mutation. The TF occupies somewhat different sites of the ESC genome in the presence of the mutation, and my goal was to determine if the motif content was distinct in the differentially bound regions. These results were published in Myers et al. (2016).

My fourth application of motifDiverge compared motif content of limb regulatory elements between bats and other mammals. This work is part of a larger computational project that I led, which forms the basis for Chapter 3. It was published in Booker, Booker, Friedrich et al. (2016)[co-first-authors] and Eckalbar et al. (2016).

2.3 Comparing binding sites between species in cardiomyocytes

I first explored the use of motifDiverge to quantify motif differences between homologous sequences. For each of the 8,225 H3K27ac-marked enhancers from mouse CMs, we identified the homologous human sequence (if any) using the whole-genome, 100-way vertebrate multiple sequence alignments available from the UCSC Genome Browser (http://genome.ucsc.edu), which are based on the hg18 and mm9 genome assemblies. It is interesting to compare CM gene regulation between these two species, because there are a number of structural and electrophysiological differences between their hearts.

I identified 1,345 orthologous human-mouse sequence pairs that were at least 20 nucleotides long. For each enhancer pair, I predicted motifs in the human and mouse sequence with JASPAR PSPMs (http://jaspar.genereg.net) for all 34 TFs expressed in mouse CMs (fragments per kilobase per million sequenced (FPKM) > 10) and a log odds score threshold that corresponds to a Type I error rate of 1%. Then I tested for TFs with significant differences in motif counts between human and mouse in each CM enhancer region.

After adjusting for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) controlling procedure (Benjamini & Hochberg, 1995), I found that most enhancers (74%) show evidence of significant differences in motif counts for at least one TF (FDR< 5%). Slightly more than half of CM enhancers (55%) have significant differences in

motif counts for multiple TFs, and several have significant differences for fifteen or more TFs. Conversely, most TFs only have significant differences in counts between human and mouse for a small percentage of CM enhancers. The TFs with the largest percentage of enhancers showing significant differences are listed in Table 2.1. These TFs are promising candidates for understanding differences in CM gene regulation between humans and mice. Interestingly, Sp1 has many enhancers with significantly more motifs in human (19%) and nearly as many with more motifs in mouse (15%), suggesting that it may target quite different sets of enhancers–and potentially different genes–in the two species.

Table 2.1 Transcription factors with the most enhancers showing significant divergence in motif counts between human and mouse sequences.

Transcription factors with more motifs in mouse	
TF	**Proportion of CM enhancers**
Prrx2	0.29
Cad	0.23
Mef2a	0.23
Arid3a	0.18
Sp1	0.15
Transcription factors with more motifs in human	
TF	**Proportion of CM enhancers**
Sp1	0.19
Egr1	0.19
Btd	0.12
Fhl1	0.083
Id1	0.080

2.4 Comparing binding sites in different cell types

Next, I used motifDiverge to compare motif counts between non-homologous sequence pairs. This application also illustrates how motifDiverge can be applied to perform a single test to compare two sets of sequences. I concatenated the sequences of the 10,338 H3K27ac-marked regions in CMs to create a single, long sequence containing all the active enhancers for this cell type. Then, I generated a similar concatenation of all 7,162 enhancers from ESCs. Any genome sequence marked by H3K27ac in both ESCs and CMs was removed from both data sets, so that the resulting two ESC and CM enhancer sequences were non-overlapping. I predicted motifs in the ESC and CM sequences as described above with PSPMs for all 49 TFs expressed in either cell type. Then I tested for TFs with significant differences in motif counts between the combined enhancer regions of the two cell types.

I found several TFs with significantly different numbers of motifs in ESC versus CM enhancers (Table 2.2 FDR< 5%). To better understand the biological meaning of these results, I used RNA-seq data from these two cell types to quantify the expression of each TF. Several TFs are only highly expressed in one cell type. For example, motif count and expression are sometimes both elevated in one cell type compared to the other. For instance, Cad is more highly expressed and has significantly more motifs in ESCs, suggesting a possibly important role in pluripotency. In other cases, such as Ctcf and Rest, the TF is expressed in both cell types, but at a lower level in the one with more motifs. For these TFs, the larger number of motifs in one cell type may be necessary to compensate for their reduced expression.

Table 2.2 Transcription factors with significant differences in motif counts between ESCs and CMs.

Expression is fragments per kilobase per million fragments sequenced (FPKM).

Transcription factors with more motifs in ESC			
TF	**FDR adjusted *p*–value**	**ESC Expression**	**CM Expression**
Arid3a	< 1e-15	4.60	14.16
Cad	< 1e-15	74.56	23.44
Prrx2	2.4e-10	3.80	33.15
Id1	2.3e-9	72.81	70.79
Nkx2-5	5.2e-6	0.96	161.63
Foxd3	0.021	17.50	0.066
Transcription factors with more motifs in CM			
TF	**FDR adjusted *p*–value**	**ESC Expression**	**CM Expression**
Ctcf	< 1e-15	38.26	13.36
Egr1	< 1e-15	17.21	167.44
Esrrb	< 1e-15	105.10	0.58
Gabpa	< 1e-15	20.43	10.57
Klf4	< 1e-15	34.51	5.34
Myc	< 1e-15	20.68	2.47
Mycn	< 1e-15	136.69	11.86
Nfil3	< 1e-15	2.75	24.077
Nfkb1	< 1e-15	9.90	13.93
Nfya	< 1e-15	6.99	15.41
Pou5f1	< 1e-15	688.11	0.13
Rela	< 1e-15	10.15	17.00
Rest	< 1e-15	44.21	12.90
Rfx1	< 1e-15	13.37	7.59
Srf	< 1e-15	21.90	29.67
Stat3	< 1e-15	10.34	39.50
Tead1	< 1e-15	13.95	25.53
Ttk	< 1e-15	18.02	2.13
Yap1	< 1e-15	30.55	37.28
Zfp423	< 1e-15	13.045	2.50
Nfe2l2	< 1e-15	24.40	22.24
Fhl1	2.82e-13	30.42	36.011
Pbx1	9.61e-11	3.33	22.94
E2f1	9.93e-11	21.093	5.48
Tbp	1.27e-08	19.075	6.62
Usf1	8.52e-08	30.35	19.79
Max	6.00e-05	27.013	16.61
Irf1	0.00023	20.89	4.25
Mef2a	0.00094	2.81	29.53
Sp1	0.033	22.83	15.57

Finally, RNA-seq data can help us filter out significant motif differences that are not biologically meaningful. For example, Nkx2-5 has significantly more motifs in ESC compared to CM enhancer sequences. However, Nkx2-5 is not expressed in ESCs, making it unlikely that the additional motifs affect ESC gene regulation. Similarly, Pou5f1 (also known as Oct4) has more motifs in CM enhancers but is not expressed in CMs, which make sense since this TF plays an important role in pluripotency (http://www.genecards.org).

2.5 Sox2 modification causes differences in binding between WT and mutant cells

I applied the motifDiverge method to a third application that measures subtle differences in binging for a modified TF compared to its unmodified state. SOX2 (sex determining region Y-box 2) is a transcription factor necessary for ESC self-renewal (Arnold et al., 2011; Masui et al., 2007). Precise control of SOX2 is critical for ESC maintenance, since increased or decreased expression of SOX2 interferes with self-renewal and pluripotency (Kopp, Ormsbee, Desler, & Rizzino, 2008; Masui et al., 2007) . Post-translational modifications (PTMs) of SOX2 may play a role in its regulation.

In this particular instance I analyzed differences in binding with and without a O-linked N-acetlyglucosamine (O-GlcNAc) modification in mouse ESCs (mESCs). O-GlcNAcylation is dynamic and O-GlcNAc signaling is essential for embryo viability (O'Donnell, Zachara, Hart, & Marth, 2004; Shafi et al., 2000; Yang et al., 2012) and mESC self-renewal (Jang et al., 2012) and O-GlcNAc transferase catalyzes this process.

While O-GlcNAc transferase is critical for mESC maintenance, the protein- and site-specific functions of O-GlcNAcylation in mESCs have not been fully elucidated.

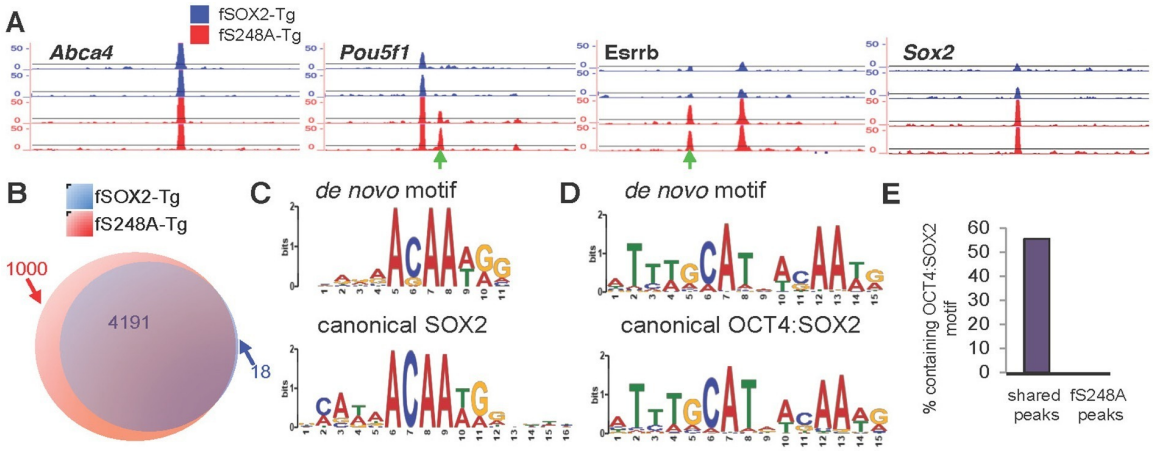
My collaborators in the Panning lab and I showed that O-GlcNAcylation of SOX2 at serine 248 (S248) is dynamically regulated in mESCs. Upon differentiation, O-GlcNAc occupancy is reduced and SOX2 is predominantly unmodified at this site. Replacement of wild type SOX2 ($SOX2^{WT}$) with an O-GlcNAc-deficient mutant SOX2 ($SOX2^{S248A}$) results in increased reprogramming efficiency. mESCs with $SOX2^{S248A}$ as their sole source of SOX2 have increased expression of genes associated with pluripotency and exhibit a decreased requirement for OCT4. $SOX2^{S248A}$ exhibits altered genomic occupancy and differential association with transcriptional regulatory complexes. Thus, our study implicates O-GlcNAc modification in coordinating genomic occupancy and protein-protein interactions of SOX2 in ESCs, and provides molecular insight into how this broadly expressed transcription factor is regulated to promote the pluripotency-specific expression program.

To examine whether the altered gene expression associated with the S248A mutation was accompanied by changes in SOX2 genomic occupancy, FLAG chromatin immunoprecipitation was performed followed by next generation sequencing (ChIP-seq) to compare SOX2 genomic distribution in fSOX2-Tg and fS248A-Tg mESCs (Figure 2.1A). SOX2 distribution exhibited considerable overlap, with 4,191 sites bound in both lines (Figure 2.1B). In addition, the mutant form of SOX2 occupied 1000 sites not bound by the wild type form (Figure 2.1A). De novo motif analysis identified the SOX2 binding

motif in fS248A-Tg specific peaks (Figure 2.1C). In mESCs, SOX2 and OCT4 heterodimerize and co-occupy a substantial portion of their target regulatory sequences (Boyer et al., 2005). De novo motif analysis of SOX2 peaks shared between fSOX2-Tg and fS248A-Tg mESCs using MEME identified the OCT4:SOX2 motif (Figure 2.1D), which was present in 2335 of the shared peaks (Bailey, Johnson, Grant, & Noble, 2015). The OCT4:SOX2 motif was not identified in any of the fS248A-Tg-specific peaks (Figure 2.1E). I found that the known OCT4:SOX2 motif was enriched in the shared peaks (FDR corrected p-value < 0.05) using motifDiverge (Kostka et al., 2014). I was able to detect this significant difference between mutant and wildtype binding despite the fact that one condition (modified) had fewer peaks. These data indicate the S248A mutation alters SOX2 genomic distribution, increasing its ability to associate with SOX2 binding sites that would not ordinarily be bound by wild type SOX2 in mESCs.

Figure 2.1 S248A mutation alters genome-wide distribution of SOX2.

(A) Representative UCSC genome browser tracks of FLAG ChIP-seq in fSOX2-Tg (blue) and fS248A-Tg (red) cells. Examples of fS248A-Tg specific peaks (Pou5f1, Esrrb) and shared peaks (Abca4, Sox2) are shown for 2 biological replicates (2 technical replicates were performed for each biological replicate, Spearman correlations for technical replicates are 1, for biological replicates 0.45 for fSOX2-Tg and 0.55 for fS248A-Tg). Each track is 15 kb. Green arrows indicate fS248A-Tg specific peaks. For Sox2 track, the region shown is not encompassed in the deletion removing endogenous Sox2. (B) Overlap (purple) in called peaks from anti-FLAG ChIP-seq in fSOX2-Tg (blue) and fS248A-Tg (red) mESCs. (C) De novo SOX2 motif identified in shared ChIP-seq peaks between fSOX2-Tg and fS248A-Tg cells (top) compared to the canonical SOX2 motif [Jaspar M01271] (bottom). (D) OCT4:SOX2 motif identified in peaks shared between fSOX2-Tg and fS248A-Tg cells using de novo motif analysis (top) compared to the canonical OCT4:SOX2 motif [Jaspar MA0142.1] (bottom). (E) Proportion of peaks containing a motif matching the OCT4:SOX2 de novo motif in shared peaks (left) and fS248A-Tg specific peaks (right).



2.6 Conclusion

These analyses show how motifDiverge can be used to analyze data from ChIP-seq experiments and how RNA-seq data can be used to filter and interpret motifDiverge findings, leading to robust conclusions about the role of sequence differences in gene regulation. I demonstrated the usefulness of comparing net changes in motif content across cell types in a differentiation time course, across species, and between mutant and wildtype cells.

3 Comparative genomics to identify limb developmental enhancers

In the previous chapter, I showed how changes in the number of binding sites for different TFs could explain differences in the regulatory potential of different cell types. I also showed that differences in binding sites between homologous regulatory regions could explain differences in species-specific gene regulation in the same cell type. Here, I attempt to go one step further and show how variation in regulatory regions may explain species-specific morphological differences.

3.1 Why limb development in bats?

The limb is a classic example of vertebrate homology and is represented by a large range of morphological structures such as fins, legs and wings. The evolution of these structures could be driven by alterations in gene regulatory elements that have critical roles during development.

The developing tetrapod limb is made up of three skeletal elements: the stylopod (humerus/femur), zeugopod (ulna/tibia, radius/fibula), and autopod (carpals/tarsals; metacarpals/metatarsals; phalanges) (Casanova & Sanz-Ezquerro, 2007; BELL, ANDRES, & GOSWAMI, 2011). Autopods are highly specialized, composed of different numbers and lengths of digits, and exhibit varying degrees of interdigital soft tissue (webbing). Autopods are a hallmark of tetrapod diversity and are essential for adaptation to life on land, in the sea and in the air. Bats are an extreme example of this. To form a wing, bat forelimbs have gone through three major changes: elongation of digits II-V,

retention of membranous tissue forming the inter-digital patagia (chiropatagium) and a relative reduction in the diameter of the ulna (L. N. Cooper & Sears, 2013; K. L. Cooper & Tabin, 2008; Sears, Behringer, Rasweiler IV, & Niswander, 2007). These morphological innovations are clearly apparent in bat fossils from 52.5 million years ago (Jepsen, 1966; Simmons, Seymour, Habersetzer, & Gunnell, 2008). The genetic changes that led to the development of these specialized limb structures and mammalian flight are likely to have occurred prior to the radiation of the Chiroptera, one of the most diverse mammalian orders.

Nucleotide changes in enhancers have previously been linked to morphological differences between species (Carroll, 2005). One such example is the *Prx1* limb enhancer. The replacement of the mouse sequence of this enhancer with the homologous bat *Prx1* sequence resulted in mice with longer forelimbs (C. J. Cretekos et al., 2008). The recent availability of several bat genomes (*Myotis lucifugus*, *Myotis davidii*, *Pteropus vampyrus*, and *Pteropus alecto*) (Zhang et al., 2013; Dong, Lei, Liu, & Zhang, 2013; Wang et al., 2014; Eckalbar et al., 2016) now make it possible to identify specific nucleotide changes in the bat lineage, as compared to other mammals, that could have a role in the development of the unique limb morphology of the bat.

3.2 Identifying enhancers controlling species specific traits

Various computational approaches have been used to identify regulatory elements that could be involved in species-specific morphological changes (Bejerano et al., 2004; Cotney et al., 2012; Dunham et al., 2012; Pollard et al., 2006; Carbone et al., 2014).

These include human accelerated regions (HARs) and human accelerated conserved noncoding sequences (HACNSs), which are highly conserved sequences that have acquired a disproportionate number of nucleotide substitutions since humans diverged from our common ancestor with chimpanzees (Pollard et al., 2006; S. Prabhakar et al., 2008; Shyam Prabhakar, Noonan, Pääbo, & Rubin, 2006). Based on epigenetic marks, my lab predicted that at least 30% of these noncoding HARs are developmental enhancers (Capra et al., 2013). So far, 62 out of 92 tested HARs have shown enhancer activity in mouse transgenic assays, and 7 out of 26 HARs, where the activity of the human and chimp sequences were compared, showed differential enhancer activity (Hubisz & Pollard, 2014). These include the limb enhancer sequences HAR2/HACNS1, which showed no limb specific activity for the non-human homologous sequence (S. Prabhakar et al., 2008), and 2xHAR.114, which displayed restricted limb activity for the human sequence compared to the chimpanzee sequence (Capra et al., 2013). These findings indicate that the identification of accelerated regions could serve to detect sequences that function as gene regulatory elements and could possibly give rise to characteristic phenotypes among species.

3.3 Computational molecular evolutionary analyses of candidate limb enhancers

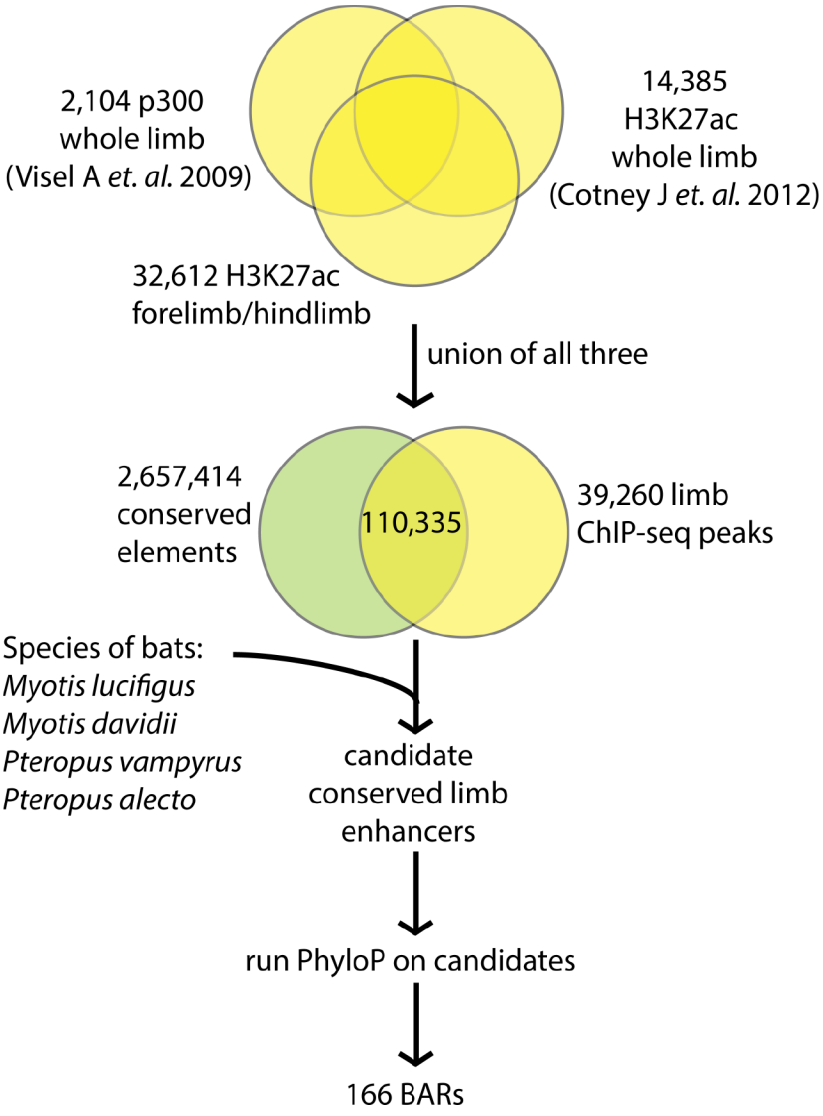
To identify BARs, I employed a statistical phylogenetic test for accelerated nucleotide evolution in the common ancestor of all extant bats. This is an extension of a previously proposed likelihood ratio test for acceleration in a single species or clade (Pollard et al., 2010). This new ancestral lineage version of the likelihood ratio test is implemented in

the PhyloP function (option—branch) in the open source software package PHAST (Hubisz, Pollard, & Siepel, 2011). The input to PhyloP is a multiple sequence alignment for each genomic region to be tested for acceleration, plus a phylogenetic tree of the species in the alignment that is estimated from genome-wide data (in this case, four-fold degenerate sites).

To apply this statistical test to bat limb development, I first identified a collection of candidate enhancers for limb development genes by intersecting evolutionarily conserved elements with enhancer-associated histone modifications and transcription factor binding events measured in the developing mouse limbs (Figure 3.1). Specifically, I took the union of all peaks from two previously published ChIP-seq experiments targeting H3K27ac or p300 (Cotney et al., 2012; Visel, Blow, et al., 2009) and an H3K27ac dataset generated for this project. Next, I generated a set of vertebrate conserved elements that were agnostic to the rate of nucleotide substitutions in bats. I started with 60-way vertebrate multiple sequence alignments with mouse as the reference species (UCSC Genome Browser, mm10 assembly). I dropped the two bat genomes (*M. lucifigus* and *P. vampyrus*) from the alignments to ensure that high rates of nucleotide differences between the bats and other vertebrates would not prevent us from identifying conservation in other species. Finally, I ran the PhastCons program with default settings (Siepel et al., 2005) on the resulting genome-wide alignments.

Figure 3.1 Computational pipeline to identify bat accelerated regions.

Limb ChIP-seq peaks were unified, then overlapped with conserved regions and then scored with PhyloP values (0 to 20) by comparing *Myotis lucifugus*, *Pteropus vampyrus*, *Myotis davidii*, and *Pteropus alecto* to 48 available vertebrate genomes. A total of 166 BAR elements were identified as accelerated regions in bats [false discovery rate (FDR) < 0.05].



This analysis identified 4,384,943 conserved elements, many of which were less than 100 bp long and, thus, too short for statistical tests for acceleration (Pollard et al., 2010). However, I observed that many short elements frequently clustered together on the chromosome and that known functional elements (e.g., coding exons) were often tiled with multiple conserved elements separated by short gaps. Hence, I iteratively merged adjacent elements until the ratio of the distance between the elements merged over the total length of the region was less than or equal to 0.1. This merging algorithm was the result of empirical experiments aimed at producing one or a small number of merged elements per exon. I also experimented with adjusting the parameters of PhastCons to produce longer elements, but found that post-processing, by merging, recapitulated exons more effectively. Next, I intersected all merged regions greater than 100 bp with the ChIP-seq peaks and unmasked the *M. lucifigus* and *P. vampyrus* sequences from the multiple alignments. Regions with more than 50% missing sequence from either bat or more than 25% of nucleotides overlapping a coding exon were dropped to produce a collection of 20,057 candidate limb enhancers.

Prior to PhyloP analysis, I integrated sequences from two additional bat genomes into the candidate enhancer alignments. I obtained assembled contigs for two bats, *M. davidii* and *P. alecto*, that were sequenced to high coverage (100x) (Zhang et al., 2013). I used the BLAST algorithm to identify alignments of the mouse sequence from each candidate enhancer to contigs from *M. davidii* and *P. alecto* (Altschul, Gish, Miller, Myers, & Lipman, 1990). The single best hit with an e-value less than or equal to 0.01 was then blasted back to the mouse genome. If this produced a reciprocal best hit (i.e., the top

scoring alignment to the mouse genome overlapped the original candidate enhancer sequence), I added the *M. davidii* or *P. alecto* sequence to the 60-way multiple alignment for that candidate enhancer. This produced alignments with between two and four bats present per enhancer. The two additional bat species were added to the phylogenetic tree corresponding to the 60-way alignments (UCSC Genome Browser) and their branch lengths were adjusted using their relationship to *M. lucifigus* and *P. vampyrus*. I then restricted our analysis to regions containing at least one bat.

Finally, I used PhyloP to test each candidate enhancer for accelerated nucleotide substitutions along the ancestral bat lineage. The resulting p-values were adjusted for multiple testing using a false discovery rate (FDR) controlling procedure (Benjamini & Hochberg, 1995; Benjamini, Drai, Elmer, Kafkafi, & Golani, 2001). I call all candidate enhancers with FDR < 5% Bat Accelerated Regions (BARs) (Table 3.1). Their genomic distribution and sequence composition were analyzed using custom Python scripts. Significant associations with functions and phenotypes of nearby genes were identified using GREAT after lifting BARs over to mm9 coordinates (McLean et al., 2010). I curated a list of limb-associated genes by exhaustively looking through the literature for evidence found in mouse or human and used resampling tests to assess associations between BARs and these genes compared to random sets of PhastCons elements.

To determine whether BARs are functional limb enhancers, we selected five BARs (BAR2, BAR4, BAR61, BAR97 and BAR116) and tested them for enhancer activity using a mouse transgenic assay. The BAR candidates were chosen based on their

location, residing within 1Mb of a known limb developmental genes whose alteration leads to a skeletal or limb phenotype (Table 3.2).

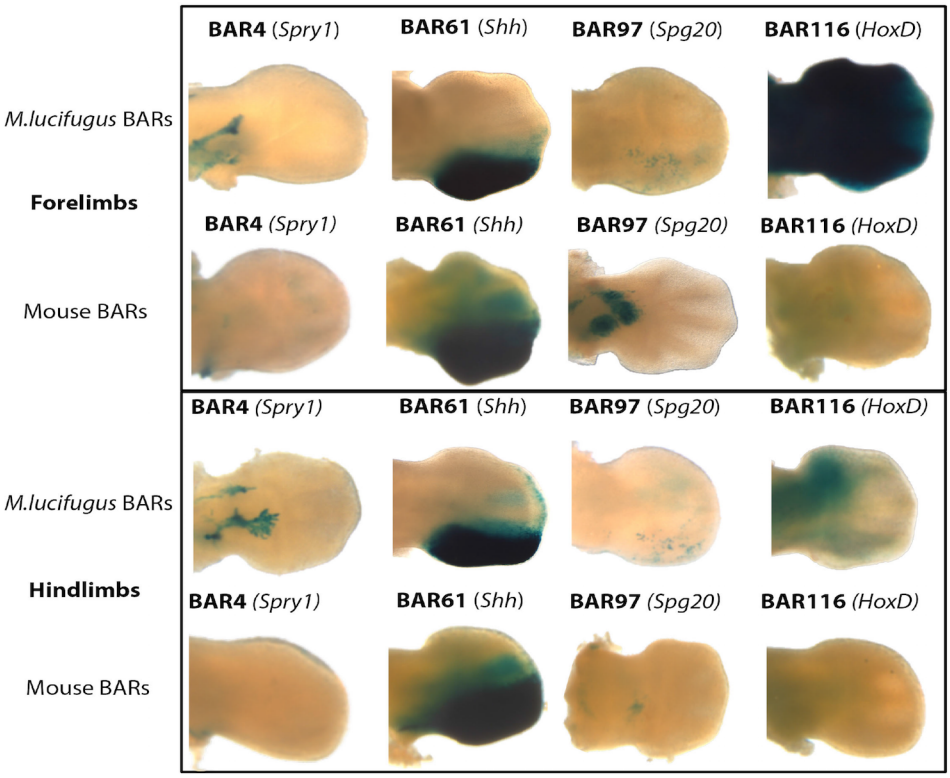
Regions spanning each of the five BAR candidate enhancers (Table 3.2; Table 3.1) were amplified from *M. lucifugus*, cloned into the Hsp68-LacZ vector that contains an *Hsp68* minimal promoter followed by the LacZ reporter gene (Kothary et al., 1988), and injected into single-cell mouse embryos. Transgenic embryos were harvested at E12.5. This stage was chosen since it is equivalent to CS16E in *Carollia perspicallata* and *Miniopterus natalensis* bat embryos, a stage when digits are identifiable and forelimbs (FL) lose their symmetry in the anterior to posterior (AP) axis compared to hindlimbs (HL) (Hockman et al., 2008; Chris J. Cretekos, Deng, Green, Rasweiler, & Behringer, 2007; Hockman, Mason, Jacobs, & Illing, 2009). All assayed *M. lucifugus* BAR sequences showed limb enhancer activity in our transgenic mouse assay (Figure 3.2).

To compare the species-specific enhancer activity of our predicted BARs, we set out to analyze the orthologous mouse sequences of four BARs (BAR4, BAR61, BAR97, BAR116; Table 3.1). Due to the nonspecific expression pattern of *M. lucifugus* BAR2, the orthologous mouse sequence was not analyzed. Regions covering each of the mouse BAR sequences were cloned into the Hsp68-LacZ vector and tested for enhancer activity at E12.5. However, the three out of the four tested mouse BAR sequences (BAR4, BAR97, BAR116) showed differential enhancer activity (Figure 3.2). Of the four, BAR4 showed differential expression in mouse compared to bat, as well as differential forelimb

and hindlimb activity. Overall, the experimental validation suggests that the accelerated sequence changes observed in BARs could lead to differences in limb enhancer expression and that my computational analysis successfully predicted these candidates.

Figure 3.2 Comparison of enhancer expression patterns for bat and mouse sequences in forelimb and hindlimb.

Representative mouse (E12.5) forelimbs (FLs) and hindlimbs (HLs) showing both *M. lucifugus* BAR and mouse BAR expression pattern. Three *M. lucifugus* BAR sequences (BAR4, 97, and 116) show differences in expression patterns as compared to the mouse BAR sequence. BAR61 (*Shh*) retains a similar expression pattern for both the bat and the mouse BAR sequences. Nearby limb-associated gene names are written in parenthesis next to the BAR ID.



3.4 TF binding site analyses of enhancers that evolved rapidly in the bat ancestor

To look for TFBS differences, I manually curated a list of limb-associated transcription

factors. BARs were analyzed for loss and gain of binding sites for each TF using motifDiverge (Kostka et al., 2014). I first compared the ancestral bat sequence to mouse. I used prequel to computationally infer the sequence of the common ancestor of extant bats using our multiple alignments (Hubisz et al., 2011). I created the corresponding aligned mouse sequence from these alignments. I then called a TFBS a hit if its FDR exceeded a threshold of 0.01. I then used motifDiverge (Kostka et al., 2014) to test if the total number of TFBS in the bat ancestor was significantly different than the number of TFBS in mouse for each TF in each individual BAR. I repeated these tests collectively over all BARs.

I next set out to identify transcription factor binding site (TFBS) changes in each of the 166 BARs by estimating the sequence of the common ancestor of the four bat genomes (*M. lucifugus*, *P. vampyrus*, *M. davidii* and *P. alecto*) and comparing this ancestral bat sequence to the orthologous mouse sequence. I predicted TFBS in the mouse and ancestral bat sequences of each BAR and tested for significant loss or gain of TFBS of 745 TFs expressed in the developing limb using motifDiverge (Kostka et al., 2014). Most TFs only had significant changes in TFBS for a single BAR, but several showed consistent patterns of loss or gain across multiple BARs. When all BARs are analyzed collectively as a single sequence, 34 TFs have significantly more TFBS in the bat ancestor compared to mouse (Table 3.4a), and 146 TFs have significantly fewer TFBS (FDR<0.05, Table 3.4b).

The most striking TFBS changes in the ancestral bat BAR sequences were gains of sites

for Nr2c2, Sp4, Zfp281, and Zfp740 each of which is enriched in twelve or more BARs. Nr2c2, also known as the testicular nuclear receptor 4 (Tr4), is involved in osteoblast maintenance and differentiation (Lin et al., 2012; Ding et al., 2013). Mice lacking Tr4 do not have apparent skeletal abnormalities, however, they display a reduction in bone mineral density and long bone volume, showing premature aging, spinal curvature (Lee et al., 2011), and osteoporosis (Lin et al., 2012). Zfp281 and Zfp740 are expressed in the developing limb (Richardson et al., 2014) but have yet to be characterized for their limb function. Two additional TFBS gains are worth noting, Egr1 and Zic2/3. The Egr genes are C2H2-type zinc finger proteins that function as transcriptional regulators with an important role in mitogenesis and differentiation. Specifically, Egr1 is involved in mouse wound repair, endochondral bone repair and data suggests that EGR1 is upregulated during skeletal muscle wound healing (Fan et al., 2013; Reumann et al., 2011). Zic2 and Zic3 belong to the C2H2-family of Zinc fingers, are known to be involved in morphogenesis and patterning during development and are associated with muscle and skeletal defects (Nagai et al., 2000; Houtmeyers, Souopgui, Tejpar, & Arkell, 2013; Garber, 1952; Quinn, Haaning, & Ware, 2012).

I also observed a significant depletion for specific TFBS when comparing the ancestral bat sequences to mice collectively over all BARs (Table 3.5a and Table 3.5b). By rank, the most depleted and fourth most depleted TFs were OSR2 and OSR1 respectively. Odd-skipped related genes, Osr1 and Osr2, belong to the C2H2 Zinc finger family (Coulter et al., 1990; Lan, Kingsley, Cho, & Jiang, 2001) and are expressed in the embryonic limb mesenchyme (So & Danielian, 1999; Stricker, Brieske, Haupt, & Mundlos, 2006). Both

Osr1 and Osr2 are associated with osteoblast regulation, chondrogenesis (Stricker et al., 2012; Verlinden et al., 2013), synovial joint formation, and their removal in mice leads to fusion of these joints (Gao, Lan, Liu, & Jiang, 2011). Also worth mentioning are Tgif1 and Meis1. Tgif1, the Thymine/Guanine interacting factor 1, is a repressor of TGF-β/Smad signaling, and is expressed in the developing limb mesenchyme (Lorda-Diez, Montero, Martinez-Cue, Garcia-Porrero, & Hurle, 2009). Meis1, a TALE homeobox TF, is a marker of the stylopod region and its overexpression abolishes distal limb structures during development (Mercader et al., 1999). Combined, our results identify TFBS gains and losses in BARs that might have a functional role.

3.5 Conclusion

Combining comparative phylogenetics with epigenetic information correctly identified four out of four enhancers as limb specific. These methods reduced the search space so that we could more accurately identify relevant enhancers. In addition, I identified an interesting forelimb specific enhancer that resides near the HoxD cluster that is expressed in bats but not mice limbs in development. I then identified motifs that might explain the functional difference between the ancestral bat and mouse sequences.

Table 3.1 BARs identified through our computational pipeline.

BAR ID	PhastCon Element (mm10)	phyloP_score	p-value	FDR
1	chr17: 12227607-12228123	20	1.00E-20	1.08E-17
2	**chr1: 91845025-91845225**	**20**	**1.00E-20**	**8.49E-18**
3	chr3: 5320701-5358677	20	1.00E-20	6.46E-18
4	**chr3: 37769126-37769766**	**20**	**1.00E-20**	**6.46E-18**
5	chr4: 17854148-17854710	20	1.00E-20	9.21E-18
6	chr7: 37338042-37338444	20	1.00E-20	1.09E-17
7	chr12: 41315104-41315629	15.955	1.11E-16	1.07E-13
8	chr7: 36977744-36979003	14.218	6.05E-15	3.31E-12
9	chr9: 35422016-35422557	13.441	3.62E-14	4.10E-11
10	chr13: 57450494-57450585	12.814	1.53E-13	1.59E-10
11	chr11: 11836728-11836993	12.775	1.68E-13	2.23E-10
12	chr15: 86366980-86367346	11.808	1.56E-12	1.51E-09
13	chr3: 8708971-8709236	11.515	3.05E-12	1.32E-09
14	chr8: 87707737-87708277	11.211	6.15E-12	5.22E-09
15	chr11: 6467579-6476087	11.089	8.15E-12	5.41E-09
16	chr18: 81602206-81602640	10.519	3.03E-11	3.32E-08
17	chr14: 21442445-21442468	10.51	3.09E-11	3.46E-08
18	chr3: 37722695-37723559	10.27	5.37E-11	1.73E-08
19	chr1: 38262359-38263461	10.032	9.29E-11	3.94E-08
20	chr9: 37146938-37147485	9.983	1.04E-10	5.89E-08
21	chr3: 8710005-8710633	9.881	1.32E-10	3.40E-08
22	chr3: 37569806-37570345	9.548	2.83E-10	6.10E-08
23	chr6: 72189017-72189223	9.385	4.12E-10	5.28E-07
24	chr18: 80554828-80555232	9.054	8.83E-10	4.84E-07
25	chr7: 70744105-70744684	8.914	1.22E-09	4.44E-07
26	chr8: 89388779-89389305	8.674	2.12E-09	8.98E-07
27	chr7: 70788912-70790131	8.492	3.22E-09	8.80E-07
28	chr3: 8865958-8866609	8.024	9.46E-09	1.75E-06
29	chr2: 28797382-28798703	7.928	1.18E-08	1.26E-05
30	chr18: 84541057-84543871	7.658	2.20E-08	8.04E-06
31	chr8: 89412095-89412621	7.431	3.71E-08	1.05E-05
32	chr7: 70625277-70625448	7.176	6.67E-08	1.46E-05

BAR ID	PhastCon Element (mm10)	phyloP_score	p-value	FDR
33	chr2: 30062653-30062676	7.108	7.80E-08	4.17E-05
34	chr12: 27502773-27502992	7.101	7.93E-08	3.84E-05
35	chr11: 12036049-12036155	7.085	8.22E-08	3.64E-05
36	chr7: 37374642-37375173	7.017	9.62E-08	1.75E-05
37	chr4: 17854015-17854076	7.014	9.68E-08	4.46E-05
38	chr13: 8871721-8871859	6.817	1.52E-07	7.89E-05
39	chr8: 89307840-89311104	6.797	1.60E-07	3.38E-05
40	chr8: 89501525-89502013	6.595	2.54E-07	4.31E-05
41	chr10: 17236031-17236080	6.581	2.62E-07	0.000264784
42	chr4: 54997477-55026531	6.394	4.04E-07	0.000123919
43	chr6: 51840057-51858080	6.338	4.59E-07	0.000294116
44	chr11: 11933009-11933203	6.296	5.06E-07	0.000167934
45	chr7: 67827353-67827643	6.229	5.90E-07	8.39E-05
46	chr7: 84109356-84110233	6.212	6.14E-07	8.39E-05
47	chr12: 40693882-40694265	6.129	7.43E-07	0.000239748
48	chrX: 58025076-58046140	6.089	8.15E-07	0.000448087
49	chr18: 77558364-77566107	6.085	8.22E-07	0.0002255
50	chr7: 66450229-66450362	6.071	8.49E-07	0.000103128
51	chr3: 102165789-102165814	6.049	8.93E-07	0.000141586
52	chr7: 70748142-70749453	6.016	9.64E-07	0.000105347
53	chr3: 41603748-41603769	6.006	9.86E-07	0.000141586
54	chr17: 84161629-84161740	5.839	1.45E-06	0.000783786
55	chr8: 87063954-87064564	5.805	1.57E-06	0.000221434
56	chr3: 8824042-8824728	5.787	1.63E-06	0.00021099
57	chr18: 83068177-83068218	5.778	1.67E-06	0.000365794
58	chr9: 41395530-41396053	5.778	1.67E-06	0.000629664
59	chr3: 9497506-9497635	5.718	1.91E-06	0.000224838
60	chr1: 38438478-38440300	5.696	2.01E-06	0.000569884
61	**chr5: 29314769-29315827**	**5.505**	**3.13E-06**	**0.002044456**
62	chr8: 87744821-87745447	5.438	3.65E-06	0.000441873
63	chr3: 55779672-55786788	5.394	4.04E-06	0.000434592
64	chr12: 5552764-5553165	5.361	4.36E-06	0.001053939
65	chr11: 36673783-36681283	5.338	4.59E-06	0.00121963
66	chr8: 89655443-89656693	5.332	4.66E-06	0.000493521
67	chr2: 27746125-27746233	5.329	4.69E-06	0.001670538
68	chr3: 42057203-42057857	5.285	5.19E-06	0.000515607
69	chr17: 35235597-35235853	5.235	5.82E-06	0.002099452
70	chr9: 41376054-41376755	5.151	7.06E-06	0.002000644
71	chr3: 37748053-37748308	5.105	7.85E-06	0.00072466
72	chr14: 56887656-56887710	5.065	8.61E-06	0.00481726
73	chr3: 104817424-104817445	5.011	9.75E-06	0.000839791
74	chr3: 9839833-9840339	4.924	1.19E-05	0.000961928

BAR ID	PhastCon Element (mm10)	phyloP_score	p-value	FDR
75	chr12: 24832460-24832484	4.884	1.31E-05	0.002528747
76	chr1: 16248957-16249903	4.847	1.42E-05	0.00260578
77	chr1: 16250230-16251318	4.814	1.53E-05	0.00260578
78	chr18: 81054253-81054374	4.725	1.88E-05	0.003443938
79	chrX: 10716430-10720712	4.715	1.93E-05	0.005300694
80	chr12: 24958932-24959853	4.68	2.09E-05	0.003370731
81	chr8: 87734811-87735028	4.65	2.24E-05	0.001983334
82	chr7: 63986554-63986819	4.632	2.33E-05	0.002145057
83	chr8: 87672172-87672303	4.631	2.34E-05	0.001983334
84	chr7: 37970018-37971554	4.628	2.36E-05	0.002145057
85	chr3: 87167821-87168059	4.618	2.41E-05	0.001831528
86	chr18: 38765976-38765997	4.616	2.42E-05	0.003794098
87	chr5: 51546693-51558460	4.606	2.48E-05	0.00810117
88	chr9: 23378237-23378892	4.575	2.66E-05	0.006029203
89	chr3: 103734206-103734294	4.534	2.92E-05	0.002098892
90	chr6: 52223075-52239016	4.516	3.05E-05	0.013014512
91	chr11: 60700264-60700288	4.465	3.43E-05	0.007586594
92	chr14: 58638375-58638691	4.446	3.58E-05	0.013356997
93	chr18: 13943061-13944213	4.441	3.62E-05	0.004967257
94	chr7: 65979097-65979188	4.417	3.83E-05	0.003218673
95	chr8: 73353213-73353283	4.396	4.02E-05	0.002878821
96	chr8: 87745553-87745726	4.39	4.07E-05	0.002878821
97	**chr3: 55527140-55527594**	**4.372**	**4.25E-05**	**0.002887413**
98	chr12: 13194635-13194660	4.316	4.83E-05	0.006680013
99	chr3: 102507418-102507508	4.268	5.40E-05	0.003485239
100	chr8: 88523666-88524537	4.25	5.62E-05	0.003668196
101	chr14: 61736734-61737320	4.248	5.65E-05	0.015804112
102	chr11: 32899695-32900137	4.168	6.79E-05	0.012885463
103	chr6: 88343544-88343855	4.144	7.18E-05	0.022987362
104	chr4: 63030328-63030487	4.13	7.41E-05	0.014597875
105	chr14: 11872799-11872828	4.121	7.57E-05	0.01693792
106	chr4: 9019315-9020181	4.101	7.93E-05	0.014597875
107	chr12: 26488507-26488748	4.047	8.97E-05	0.010177306
108	chr12: 27187517-27188012	4.024	9.46E-05	0.010177306
109	chr8: 89523281-89523652	4.018	9.59E-05	0.005811227
110	chr14: 78538648-78538668	4.014	9.68E-05	0.018058382
111	chr4: 8910656-8911073	3.988	0.000102802	0.01578005
112	chr7: 82703066-82703100	3.967	0.000107895	0.008423491
113	chr8: 88570580-88570844	3.943	0.000114025	0.006446212
114	chrX: 81071908-81071937	3.93	0.00011749	0.018167815
115	chr18: 83096839-83097471	3.92	0.000120226	0.014654268
116	**chr2: 75208968-75209651**	**3.91**	**0.000123027**	**0.032878933**

BAR ID	PhastCon Element (mm10)	phyloP_score	p-value	FDR
117	chr17: 10335120-10335606	3.906	0.000124165	0.033586695
118	chr5: 30911566-30911587	3.893	0.000127938	0.027890512
119	chrX: 10216710-10216753	3.879	0.00013213	0.018167815
120	chr1: 13139037-13142796	3.863	0.000137088	0.019397977
121	chr18: 83110490-83110957	3.845	0.000142889	0.015674967
122	chr8: 87860931-87861869	3.821	0.000151008	0.007672677
123	chr8: 89383239-89383441	3.813	0.000153815	0.007672677
124	chr4: 58677747-58677772	3.795	0.000160325	0.021094129
125	chr11: 6000677-6000724	3.793	0.000161065	0.024319552
126	chr11: 76477183-76477341	3.783	0.000164816	0.024319552
127	chr8: 89387412-89388141	3.757	0.000174985	0.007936742
128	chr8: 70905892-70905962	3.75	0.000177828	0.007936742
129	chr6: 98690273-98694473	3.742	0.000181134	0.046406533
130	chr12: 25099981-25100133	3.737	0.000183231	0.017736804
131	chr17: 5233498-5233593	3.681	0.000208449	0.045108383
132	chr7: 25267402-25276273	3.673	0.000212324	0.015471375
133	chr9: 57639351-57639378	3.668	0.000214783	0.040558199
134	chrX: 36988732-36988875	3.636	0.000231206	0.025432713
135	chr14: 70766806-70766859	3.628	0.000235505	0.037647145
136	chr3: 86777334-86777744	3.581	0.000262422	0.016145192
137	chr7: 66933951-66933974	3.579	0.000263633	0.018009439
138	chr9: 13517946-13518237	3.559	0.000276058	0.044681924
139	chr8: 89721566-89722084	3.551	0.00028119	0.01192246
140	chr3: 5200892-5205193	3.502	0.000314775	0.018485867
141	chr7: 6156115-6156245	3.484	0.000328095	0.021006208
142	chr12: 69494749-69494775	3.479	0.000331894	0.029206712
143	chr4: 14273368-14274166	3.471	0.000338065	0.038919714
144	chr9: 88521825-88521862	3.467	0.000341193	0.048321446
145	chr7: 90129228-90129627	3.461	0.000345939	0.021006208
146	chr7: 72215502-72216324	3.43	0.000371535	0.021373053
147	chr8: 89107913-89108450	3.421	0.000379315	0.0153171
148	chr3: 37666746-37667178	3.374	0.000422669	0.02374295
149	chr8: 48308821-48309203	3.323	0.000475335	0.018322012
150	chr7: 100918274-100918412	3.299	0.000502343	0.027453023
151	chr3: 87910037-87910069	3.276	0.000529663	0.028513549
152	chr8: 96488767-96490407	3.191	0.000644169	0.023750241
153	chr7: 65803863-65803893	3.162	0.000688652	0.035842712
154	chr8: 90876236-90876744	3.145	0.000716143	0.025303734
155	chr8: 87691175-87691204	3.077	0.000837529	0.028408993
156	chr7: 70705855-70706632	3.061	0.00086896	0.043171534
157	chr8: 87794695-87795196	3.033	0.00092683	0.030228911
158	chr7: 19320007-19320039	3.02	0.000954993	0.045382909

BAR ID	PhastCon Element (mm10)	phyloP_score	p-value	FDR
159	chr8: 102983474-102983496	2.938	0.001153453	0.035746854
160	chr8: 89047320-89047467	2.928	0.001180321	0.035746854
161	chr8: 77350658-77350684	2.844	0.001432188	0.041330875
162	chr8: 87690839-87690865	2.835	0.001462177	0.041330875
163	chr8: 89383498-89383761	2.807	0.001559553	0.042661307
164	chr8: 87151890-87152756	2.773	0.001686553	0.044693655
165	chr8: 87690898-87690949	2.756	0.001753881	0.045069414
166	chr8: 91444731-91445344	2.706	0.001967886	0.049081399

Table 3.2 BARs selected for mouse enhancer assays.

BARs that were selected for enhancer assays, the limb-associated genes nearby, the limb phenotype caused by mutations in these genes.

BAR ID	Nearby Limb Genes	Limb-Associated Phenotypes (MGI, OMIM) & Tissue Expression
2	*Twist2*	Skeletal and muscle abnormalities
4	*Spry1*	Chondrodysplasia, muscles, tendons
61	*Shh*	Limb malformations
97	*Spg20*	Spastic paraplegias
116	*HoxD* cluster	Skeletal defects

Table 3.3a The number of limb-associated transcription factors with significant binding site gains summed up across all BARs.

Number of BARs that show enrichment (FDR < 0.05)	Transcription Factor
1	ALX3
1	ARX
1	BARX1
1	BARX2
2	E2F2
2	E2F3
1	EGR1
1	EGR2
5	GLIS2
1	HOXA5
1	KLF7
1	LHX1
1	LHX6
1	LMX1B
1	MEF2A
12	NR2C2
1	PAX7
6	PLAGL1
1	PRRX2
1	SOX21
12	SP4
1	TBP
4	TCFAP2A
2	TCFAP2B
3	TCFAP2C
3	TCFAP2E
5	ZBTB7B
2	ZFP105
5	ZFP161
12	ZFP281
13	ZFP740
4	ZIC1
5	ZIC2
4	ZIC3

Table 3.3b The number of limb-associated transcription factors with significant binding site losses summed up across all BARs.

Number of BARs that show depletion (FDR < 0.05)	Transcription Factor
1	ALX3
1	ARID5A
3	ASCL2
2	ATF1
1	BARHL1
1	BARX1
2	BARX2
1	BBX
1	BCL6
3	BCL6B
1	BSX
1	CDX1
2	CPHX
2	CUTL1
3	DBX1
2	DBX2
1	DLX1
1	DLX2
2	DLX3
1	DLX4
2	DLX5
1	E2F3
3	EGR2
1	EHF
1	EMX2
1	EN2
2	ESR2
1	ESRRA
1	ESRRB
1	EVX1
1	EVX2
2	GBX1
2	GBX2
3	GM397
1	HBP1
1	HIST1H2BN
1	HMBOX1

Number of BARs that show depletion (FDR < 0.05)	Transcription Factor
1	HMX1
1	HMX2
1	HNF1A
2	HNF4A
1	HNF4G
1	HOXA11
1	HOXA2
1	HOXA4
1	HOXA5
3	HOXA7
1	HOXB3
1	HOXB5
1	HOXB6
2	HOXB7
1	HOXB8
1	HOXC10
1	HOXC11
1	HOXC5
2	HOXC8
1	HOXD1
1	HOXD11
1	HOXD3
2	HOXD8
2	IRX2
2	IRX3
2	IRX4
2	IRX5
2	IRX6
1	ISGF3G
1	ISL2
1	ISX
2	JUNDM2
1	LBX2
1	LEF1
1	LHX2
1	LHX4
1	LHX6
2	LHX8
1	LMX1A
1	LMX1B
2	MAFB
1	MAFF

Number of BARs that show depletion (FDR < 0.05)	Transcription Factor
2	MAFK
2	MEIS1
2	MEOX1
2	MRG2
1	MSX1
1	MSX2
1	MYBL1
3	MYF6
2	NFIC
1	NFYA
1	NHLH1
1	NKX1-2
2	NKX2-2
2	NKX2-3
1	NKX2-5
1	NKX2-9
2	NKX3-1
2	NKX6-1
1	NKX6-3
1	NR2F1
1	NR2F2
1	OSR1
3	OSR2
1	PBX1
2	PKNOX1
2	PKNOX2
1	POU1F1
1	POU2F1
1	POU2F2
1	POU3F1
1	POU3F2
1	POU3F4
1	PRRX2
2	RARA
1	RFX4
4	RHOX11
1	RHOX6
1	RXRA
1	SFPI1
1	SIX1
1	SIX3
2	SIX6

Number of BARs that show depletion (FDR < 0.05)	Transcription Factor
1	SMAD3
1	SOX1
1	SOX12
1	SOX13
1	SOX15
1	SOX18
1	SOX21
1	SOX30
1	SOX5
2	STAT1
1	STAT3
1	STAT4
2	TCFCP2L1
2	TGIF1
2	TGIF2
1	TITF1
3	TLX2
1	VAX1
1	VAX2
1	ZBTB12
3	ZBTB3
2	ZFP105
2	ZFP161
1	ZFP187
3	ZFP691

Table 3.4a Limb-associated transcription factors with significant (FDR < 0.05) gains in binding sites in all BARs collectively.

Transcription Factor	# Ancestral bat TFBS	# M. musculus (mm10) TFBS	FDR for TFBS gain in ancestral bat
NR2C2	23357	429	0
SP4	23114	158	0
ZFP281	22918	27	0
ZFP740	23588	84	0
GLIS2	1151	481	9.12E-43
ZBTB7B	728	286	1.62E-30
ZIC1	436	137	4.11E-26
ZIC3	593	228	6.10E-26
PAX4	666	272	6.59E-26
ZIC2	622	249	2.62E-25
PLAGL1	685	309	1.53E-21
EGR1	242	69	1.66E-16
KLF7	473	262	5.09E-08
E2F3	871	594	5.13E-05
ZFP161	693	466	0.000205709
E2F2	767	532	0.000667945
TCFAP2A	2129	1687	0.032926745
ZFX	413	287	0.049940341

Table 3.4b Limb-associated transcription factors with significant (FDR < 0.05) losses in binding sites in all BARs collectively.

Transcription Factor	# Ancestral bat TFBS	# M. musculus (mm10) TFBS	FDR for TFBS loss in ancestral bat
OSR2	3206	3496	6.33E-17
GM397	4701	4908	5.04E-16
TGIF1	4561	4707	1.74E-13
OSR1	3090	3273	2.54E-12
SMAD3	2288	2487	6.04E-12
TGIF2	4166	4254	8.44E-11
MEIS1	4372	4429	3.36E-10
PKNOX2	4308	4358	6.93E-10
MRG2	4298	4346	7.55E-10
PKNOX1	4202	4251	9.94E-10
MYF6	4223	4256	2.82E-09
ZFP691	3201	3287	4.27E-09
MAFB	4343	4354	6.48E-09
NKX2-2	3578	3636	6.48E-09
HNF4G	3461	3499	5.07E-08
NFIC	3284	3335	5.07E-08
BCL6	2107	2201	1.91E-07
NKX2-3	3658	3654	3.92E-07
HNF4A	3524	3527	4.32E-07
DLX5	3349	3346	1.41E-06
NR2F2	3086	3097	1.67E-06
CPHX	2920	2938	2.15E-06
RFX3	2736	2762	2.23E-06
MAFK	3504	3475	3.55E-06
HOXB3	3147	3140	3.81E-06
RARA	3915	3853	4.11E-06
HSF1	1872	1938	4.12E-06
ESRRB	3430	3401	4.46E-06
EVX2	3478	3444	4.46E-06
EVX1	2939	2940	4.75E-06
ATF1	3200	3182	5.08E-06
ESR2	3889	3822	5.08E-06
GABPA	2041	2093	5.08E-06
STAT1	2362	2398	5.08E-06
RFX4	2033	2085	5.13E-06
TITF1	2734	2745	5.13E-06
ASCL2	4168	4073	6.82E-06
TP63	1466	1538	8.60E-06

Transcription Factor	# Ancestral bat TFBS	# M. musculus (mm10) TFBS	FDR for TFBS loss in ancestral bat
ZBTB3	2973	2959	9.03E-06
MAFF	2185	2218	1.05E-05
RFXDC2	1894	1942	1.17E-05
HMBOX1	2611	2616	1.18E-05
EMX2	3255	3215	1.27E-05
BCL6B	3900	3806	1.89E-05
ZBTB12	2760	2747	1.98E-05
GBX2	3131	3085	3.13E-05
LHX8	3728	3634	3.80E-05
IRX3	2311	2314	4.85E-05
ESRRA	3908	3795	4.87E-05
RHOX11	4017	3893	5.45E-05
BSX	3308	3238	5.74E-05
NKX2-9	3186	3120	7.58E-05
AR	435	506	0.000119551
MEOX1	3123	3052	0.000134301
SOX9	1098	1154	0.000135528
LEF1	2479	2453	0.000143288
NKX3-1	2489	2461	0.000153532
NKX2-5	2963	2897	0.000191711
NKX3-1	3135	3055	0.000196032
IRF6	1612	1636	0.000201225
MEIS1	523	588	0.000239871
STAT3	1632	1653	0.000241158
ISX	3407	3300	0.00025902
SIX1	2014	2004	0.000366978
IRX3	2101	2084	0.000390485
BHLHB2	3332	3220	0.000413714
NFYA	1959	1950	0.000417398
HOXD3	2721	2655	0.000484818
STAT6	1395	1419	0.000484818
SOX17	373	432	0.000501868
JUNDM2	3342	3226	0.000502801
STAT4	1553	1566	0.000537459
CUTL1	3790	3632	0.000620996
IRX5	2226	2188	0.000766494
RHOX11	4055	3866	0.000908046
SFPI1	1357	1374	0.000920502
IRX4	1851	1835	0.000986665
EHF	2216	2174	0.001004256
GBX1	2090	2056	0.001041848

Transcription Factor	# Ancestral bat TFBS	# M. musculus (mm10) TFBS	FDR for TFBS loss in ancestral bat
HBP1	2108	2072	0.001088512
DLX2	2534	2465	0.0011421
SPDEF	2900	2802	0.0011421
ESR1	2318	2262	0.001316075
NKX2-6	1984	1952	0.001405533
ZFP187	1703	1691	0.001405533
SIX4	2450	2382	0.001471647
MSX2	2957	2843	0.0019507
IRF4	1600	1588	0.002158649
IRX6	1807	1780	0.002181153
SOX12	946	973	0.002190947
LHX4	2612	2522	0.002333893
TCFCP2L1	2805	2699	0.002432326
NR5A2	1269	1275	0.002622653
PAX7	2410	2332	0.002847715
NKX2-4	2289	2220	0.00291678
NR4A1	381	425	0.00291678
RFX2	1898	1856	0.003329914
HMX1	2510	2419	0.003545841
SRF	1962	1914	0.003545841
CEBPG	256	298	0.003777847
NKX1-2	2827	2707	0.0039201
SIX6	2875	2751	0.0039201
VAX2	2760	2646	0.0039201
RPP25	466	505	0.004028288
EOMES	2059	2000	0.004164052
RXRA	2680	2570	0.004295331
ZDHHC15	461	498	0.004969429
POU6F1	2743	2624	0.005080929
HOXB5	2442	2348	0.005139827
POU2F2	2570	2465	0.005146892
REST	1610	1580	0.005146892
SRF	3493	3308	0.005211699
RUNX1	459	495	0.005252133
IRF5	1184	1182	0.005769343
SOX5	1109	1111	0.006018737
PPARG	925	937	0.006351575
SOX1	1631	1595	0.006699239
LHX6	3336	3157	0.006814996
IRX2	2049	1979	0.007167913
ELF3	1481	1454	0.007603498

Transcription Factor	# Ancestral bat TFBS	# M. musculus (mm10) TFBS	FDR for TFBS loss in ancestral bat
CRX	2479	2372	0.007710044
CBFB	1169	1164	0.00772033
PITX2	1898	1838	0.007769352
SIX6	2756	2624	0.00790482
MAFK	697	718	0.008017415
PRRX2	2281	2189	0.008017415
MAFB	883	894	0.008257736
PAX2	3088	2927	0.008413564
HOXA7	2567	2449	0.008563017
ISGF3G	912	920	0.008563017
HOXD1	3057	2895	0.008772778
BBX	2013	1940	0.009099944
PPP5C	404	435	0.009512204
SOX7	1312	1291	0.009512204
EN1	1150	1141	0.009693508
SIX2	2650	2520	0.010159901
POU6F1	2181	2090	0.010663258
PRRX1	1801	1741	0.010663258
SOX4	1134	1124	0.010663258
ABCF2	181	213	0.010905424
SIX3	2817	2669	0.011387538
TCF3	2519	2397	0.011449184
ARX	2180	2086	0.012001372
SHOX2	2490	2368	0.012730284
GATA5	1529	1486	0.013088481
HSF1	695	709	0.013443192
ZFP128	2149	2053	0.014735692
HOXA2	2706	2561	0.014962559
HOXA5	798	804	0.01558195
SOX30	2111	2016	0.016291486
BARX1	2733	2583	0.016440163
LBX2	2628	2487	0.016486561
MYOD1	262	290	0.016900787
PBX1	1662	1603	0.016900787
SIX6	2698	2548	0.018217478
EN2	2573	2434	0.018315207
ESRRA	913	908	0.019146346
LHX2	2290	2175	0.019146346
IRF1	268	294	0.020141001
MAP4K2	180	207	0.020141001
PAXIP1	189	216	0.020141001

Transcription Factor	# Ancestral bat TFBS	# M. musculus (mm10) TFBS	FDR for TFBS loss in ancestral bat
POU2F3	1708	1641	0.020141001
VAX1	2836	2670	0.020141001
ZFP410	1117	1096	0.020141001
HOXC11	2073	1975	0.020155977
FOXJ3	911	904	0.020528527
NR2F1	296	321	0.020678245
PHOX2A	2446	2314	0.020892588
LHX6	2427	2296	0.021238926
PAX4	2392	2263	0.022220351
HOXA3	2450	2316	0.022489766
HOXB6	2629	2478	0.022710119
MYB	2770	2605	0.023305562
PICK1	244	269	0.023305562
HOXA5	2590	2441	0.023864679
DLX3	2693	2534	0.023939885
HDX	2754	2589	0.023939885
ING3	218	243	0.023939885
NFE2L2	557	569	0.023939885
PAX5	1179	1149	0.023939885
NKX2-3	438	456	0.024039615
INSM1	166	191	0.024046905
TBP	311	333	0.025283229
FOXC1	236	260	0.025344035
GSC	2040	1937	0.025757604
HOXC13	2154	2041	0.025757604
MRPL1	335	356	0.025757604
RPS4X	130	154	0.025757604
RBM8A	266	289	0.026001693
PAX6	1303	1261	0.026664017
FOXJ1	937	922	0.027558091
HOXA11	2092	1981	0.02957612
DBX1	682	683	0.029609452
HOXA13	2582	2426	0.029813174
SOX18	1526	1463	0.030459996
EN1	1942	1843	0.030752296
POU3F2	1188	1151	0.032386409
HIC1	2237	2110	0.032724036
MAX	2674	2506	0.033401635
RAX	2204	2079	0.033569318
LARP4	521	530	0.033572534
PAX6	871	857	0.033791921

Transcription Factor	# Ancestral bat TFBS	# M. musculus (mm10) TFBS	FDR for TFBS loss in ancestral bat
DLX4	2787	2607	0.034053898
HOXB7	1965	1861	0.034053898
VAX2	451	463	0.034989094
RHOX6	2060	1946	0.035663517
GATA6	1685	1604	0.035797824
HOXC12	2640	2472	0.035982766
VSX1	1711	1627	0.036761383
MXD4	153	174	0.037750821
OTX2	2196	2064	0.044349326
APEX2	244	262	0.044768215
PITX3	1905	1799	0.0454132
PHOX2B	1671	1585	0.046762879
SPIB	284	300	0.047584857
HOXD12	2637	2460	0.049454816
MSI1	139	158	0.049454816
ALX3	2361	2210	0.049721234
HNF4A	3208	2975	0.049759952

4 Summary

During my graduate career, I was interested in how genetic variation affects phenotype through gene regulation. Using bioinformatic tools that identify changes at the sequence level, I was able to identify sequence changes between species and between cell types that potentially explain functional differences between the conditions I was comparing. In Chapter 2, I identified transcription factors that showed a significant difference in the number of motifs enriched in homologous mouse and human cardiomyocyte enhancers. I also did a similar comparison between enhancers found in embryonic stem cells and differentiated cardiomyocytes. These motif differences could characterize the functional differences between enhancers found in cardiomyocytes. I also applied these methods to a third dataset and was able to show depletion of the OCT4:SOX2 motif in SOX2 peaks where SOX2 contained the S248A mutation. In Chapter 3, I look at sequence variation across mammals and ask if there exist an unusually higher than expected number of substitutions along the bat lineage than expected by chance. With my collaborators, I was able to test these candidate enhancers using transgenic *in vivo* reporter assays and correctly identify limb specific enhancers. In particular, I helped point to the discovery of an enhancer near the HoxD cluster that shows forelimb specific expression in bats compared to mice. Using various computational methods, I quantified differences between enhancers using motifs and also prioritized enhancers in order to identify ones that were functionally relevant within a particular context.

References

Alföldi, J., & Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. *Genome Research*. https://doi.org/10.1101/gr.157503.113

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2

An integrated encyclopedia of DNA elements in the human genome. (2012). *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Arnold, K., Sarkar, A., Yram, M. A., Polo, J. M., Bronson, R., Sengupta, S., … Hochedlinger, K. (2011). Sox2 + adult stem and progenitor cells are important for tissue regeneration and survival of mice. *Cell Stem Cell*, *9*(4), 317–329. https://doi.org/10.1016/j.stem.2011.09.001

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, *43*(W1), W39–W49. https://doi.org/10.1093/nar/gkv416

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., … Zheng, Y. T. (2010). Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders. *Proceedings of the National Academy of Sciences*, *107*(25), 11459–11464. https://doi.org/10.1073/pnas.1002443107

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science (New York, N.Y.)*, *304*(5675), 1321–1325. https://doi.org/10.1126/science.1098119

BELL, E., ANDRES, B., & GOSWAMI, A. (2011). Integration and dissociation of limb elements in flying vertebrates: a comparison of pterosaurs, birds and bats. *Journal of Evolutionary Biology*, *24*(12), 2586–2599. https://doi.org/10.1111/j.1420-9101.2011.02381.x

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research*, *125*(1–2), 279–284. https://doi.org/10.1016/S0166-4328(01)00297-2

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., … Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state

identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, *44*(2), 148–156. https://doi.org/10.1038/ng.1064

Booker, B. M., Friedrich, T., Mason, M. K., VanderMeer, J. E., Zhao, J., Eckalbar, W. L., … Ahituv, N. (2016). Bat Accelerated Regions Identify a Bat Forelimb Specific Enhancer in the HoxD Locus. *PLoS Genetics*, *12*(3). https://doi.org/10.1371/journal.pgen.1005738

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., … Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, *122*(6), 947–956. https://doi.org/10.1016/j.cell.2005.08.020

Bradley, R. K., Li, X. Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H. C., … Eisen, M. B. (2010). Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related drosophila species. *PLoS Biology*, *8*(3). https://doi.org/10.1371/journal.pbio.1000343

Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R., Pollard, K. S., Jiang, Z., … Matys, V. (2013). Many human accelerated regions are developmental enhancers. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1632), 20130025. https://doi.org/10.1098/rstb.2013.0025

Carbone, L., Alan Harris, R., Gnerre, S., Veeramah, K. R., Lorente-Galdos, B., Huddleston, J., … Gibbs, R. A. (2014). Gibbon genome and the fast karyotype evolution of small apes. *Nature*, *513*(7517), 195–201. https://doi.org/10.1038/nature13679

Carroll, S. B. (2005). Evolution at Two Levels: On Genes and Form. *PLoS Biology*, *3*(7), e245. https://doi.org/10.1371/journal.pbio.0030245

Casanova, J. C., & Sanz-Ezquerro, J. J. (2007). Digit morphogenesis: Is the tip different? *Development, Growth & Differentiation*, *49*(6), 479–491. https://doi.org/10.1111/j.1440-169X.2007.00951.x

Consortium, T. Gte., Ardlie, K. G., Deluca, D. S., Segrè, A. V., Sullivan, T. J., Young, T. R., … Dermitzakis, E. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Cooper, K. L., & Tabin, C. J. (2008). Understanding of bat wing evolution takes flight. *Genes & Development*, *22*(2), 121–124. https://doi.org/10.1101/gad.1639108

Cooper, L. N., & Sears, K. E. (2013). How to Grow a Bat Wing. In *Bat Evolution, Ecology, and Conservation* (pp. 3–20). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7397-8_1

Cotney, J., Leng, J., Oh, S., DeMare, L. E., Reilly, S. K., Gerstein, M. B., & Noonan, J. P. (2012). Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Research*, *22*(6), 1069–1080. https://doi.org/10.1101/gr.129817.111

Coulter, D. E., Swaykus, E. A., Beran-Koehn, M. A., Goldberg, D., Wieschaus, E., & Schedl, P. (1990). Molecular analysis of odd-skipped , a zinc finger encoding segmentation gene with a novel pair-rule expression pattern. *Embo J.*, *8*(12), 3795–3804. https://doi.org/2120051

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., … Collins, F. S. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, *16*(1), 123–131. https://doi.org/10.1101/gr.4074106

Cretekos, C. J., Deng, J. M., Green, E. D., Rasweiler, J. J., & Behringer, R. R. (2007). Isolation, genomic structure and developmental expression of Fgf8 in the short-tailed fruit bat, Carollia perspicillata. *International Journal of Developmental Biology*, *51*(4), 333–338. https://doi.org/10.1387/ijdb.062257cc

Cretekos, C. J., Wang, Y., Green, E. D., Martin, J. F., Rasweiler, J. J., & Behringer, R. R. (2008). Regulatory divergence modifies limb length between mammals. *Genes & Development*, *22*(2), 141–151. https://doi.org/10.1101/gad.1620408

Ding, X., Yu, S., Chen, B., Lin, S., Chang, C., & Li, G. (2013). Recent advances in the study of testicular nuclear receptor 4. *Journal of Zhejiang University. Science. B*, *14*(3), 171–7. https://doi.org/10.1631/jzus.B1200357

Dong, D., Lei, M., Liu, Y., & Zhang, S. (2013). Comparative inner ear transcriptome analysis between the Rickett's big-footed bats (Myotis ricketti) and the greater short-nosed fruit bats (Cynopterus sphinx). *BMC Genomics*, *14*, 916. https://doi.org/10.1186/1471-2164-14-916

Eckalbar, W. L., Schlebusch, S. A., Mason, M. K., Gill, Z., Parker, A. V, Booker, B. M., … Ahituv, N. (2016). Transcriptomic and epigenomic characterization of the developing bat wing. *Nature Genetics*, *48*(5), 528–536. https://doi.org/10.1038/ng.3537

Fan, Y. Y., Ye, G. H., Lin, K. Z., Yu, L. S., Wu, S. Z., Dong, M. W., … Li, X. B. (2013). Time-dependent expression and distribution of Egr-1 during skeletal muscle wound healing in rats. *Journal of Molecular Histology*, *44*(1), 75–81. https://doi.org/10.1007/s10735-012-9445-8

Gao, Y., Lan, Y., Liu, H., & Jiang, R. (2011). The zinc finger transcription factors Osr1 and Osr2 control synovial joint formation. *Developmental Biology*, *352*(1), 83–91. https://doi.org/10.1016/j.ydbio.2011.01.018

Harmston, N., Baresic, A., & Lenhard, B. (2013). The mystery of extreme non-coding conservation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1632), 20130021. https://doi.org/10.1098/rstb.2013.0021

Hockman, D., Cretekos, C. J., Mason, M. K., Behringer, R. R., Jacobs, D. S., & Illing, N. (2008). A second wave of Sonic hedgehog expression during the development of the bat limb. *Proceedings of the National Academy of Sciences*, *105*(44), 16982–16987. https://doi.org/10.1073/pnas.0805308105

Hockman, D., Mason, M. K., Jacobs, D. S., & Illing, N. (2009). The role of early development in mammalian limb diversification: A descriptive comparison of early limb development between the natal long-fingered bat (miniopterus natalensis) and the mouse (mus musculus). *Developmental Dynamics*, *238*(4), 965–979. https://doi.org/10.1002/dvdy.21896

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., … Noble, W. S. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, *41*(2), 827–841. https://doi.org/10.1093/nar/gks1284

Houtmeyers, R., Souopgui, J., Tejpar, S., & Arkell, R. (2013). The ZIC gene family encodes multi-functional proteins essential for patterning and morphogenesis. *Cellular and Molecular Life Sciences*. https://doi.org/10.1007/s00018-013-1285-5

Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2014.07.005

Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). Phastand Rphast: Phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, *12*(1), 41–51. https://doi.org/10.1093/bib/bbq072

Jang, H., Kim, T. W., Yoon, S., Choi, S. Y., Kang, T. W., Kim, S. Y., … Youn, H. D. (2012). O-GlcNAc regulates pluripotency and reprogramming by directly acting on core components of the pluripotency network. *Cell Stem Cell*, *11*(1), 62–74. https://doi.org/10.1016/j.stem.2012.03.001

Jepsen, G. L. (1966). Early Eocene Bat from Wyoming. *Science*, *154*(3754), 1333–1339. https://doi.org/10.1126/science.154.3754.1333

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Kopp, J. L., Ormsbee, B. D., Desler, M., & Rizzino, A. (2008). Small Increases in the Level of Sox2 Trigger the Differentiation of Mouse Embryonic Stem Cells. *Stem*

Cells, *26*(4), 903–911. https://doi.org/10.1634/stemcells.2007-0951

Kostka, D., Friedrich, T., Holloway, A. K., & Pollard, K. S. (2014). motifDiverge: a model for assessing the statistical significance of gene regulatory motif divergence between two DNA sequences. Genomics. Retrieved from http://arxiv.org/abs/1402.0042

Kothary, R., Clapoff, S., Brown, A., Campbell, R., Peterson, A., & Rossant, J. (1988). A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube. *Nature*. https://doi.org/10.1038/335435a0

Kvon, E. Z. (2015). Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics*. https://doi.org/10.1016/j.ygeno.2015.06.007

Lan, Y., Kingsley, P. D., Cho, E. S., & Jiang, R. (2001). Osr2, a new mouse gene related to Drosophila odd-skipped, exhibits dynamic expression patterns during craniofacial, limb, and kidney development. *Mechanisms of Development*, *107*(1–2), 175–179. https://doi.org/10.1016/S0925-4773(01)00457-9

Lee, Y.-F., Liu, S., Liu, N.-C., Wang, R.-S., Chen, L.-M., Lin, W.-J., … Chang, C. (2011). Premature aging with impaired oxidative stress defense in mice lacking TR4. *American Journal of Physiology. Endocrinology and Metabolism*, *301*(1), E91-8. https://doi.org/10.1152/ajpendo.00701.2010

Lin, S.-J., Ho, H.-C., Lee, Y.-F., Liu, N.-C., Liu, S., Li, G., … Chang, C. (2012). Reduced osteoblast activity in the mice lacking TR4 nuclear receptor leads to osteoporosis. *Reproductive Biology and Endocrinology : RB&E*, *10*, 43. https://doi.org/10.1186/1477-7827-10-43

Lorda-Diez, C. I., Montero, J. A., Martinez-Cue, C., Garcia-Porrero, J. A., & Hurle, J. M. (2009). Transforming growth factors ?? coordinate cartilage and tendon differentiation in the developing limb mesenchyme. *Journal of Biological Chemistry*, *284*(43), 29988–29996. https://doi.org/10.1074/jbc.M109.014811

Maston, G. a, Landt, S. G., Snyder, M., & Green, M. R. (2012). Characterization of enhancer function from genome-wide analyses. *Annual Review of Genomics and Human Genetics*, *13*(1), 29–57. https://doi.org/10.1146/annurev-genom-090711-163723

Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., … Niwa, H. (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol*, *9*(6), 625-U26. https://doi.org/Doi 10.1038/Ncb1589

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., … Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory

regions. *Nature Biotechnology*, *28*(5), 495–501. https://doi.org/10.1038/nbt.1630

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., … Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, *30*(3), 271–7. https://doi.org/10.1038/nbt.2137

Mercader, N., Leonardo, E., Azpiazu, N., Serrano, A, Morata, G., Martínez, C., & Torres, M. (1999). Conserved regulation of proximodistal limb axis development by Meis1/Hth. *Nature*, *402*(6760), 425–9. https://doi.org/10.1038/46580

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., … Duboule, D. (2011). A Regulatory Archipelago Controls Hox Genes Transcription in Digits. *Cell*, *147*(5), 1132–1145. https://doi.org/10.1016/j.cell.2011.10.023

Myers, S. A., Peddada, S., Chatterjee, N., Friedrich, T., Tomoda, K., Krings, G., … Panning, B. (2016). SOX2 O-GlcNAcylation alters its protein-protein interactions and genomic occupancy to modulate gene expression in pluripotent cells. *eLife*, *5*(MARCH2016). https://doi.org/10.7554/eLife.10647

Nagai, T., Aruga, J., Minowa, O., Sugimoto, T., Ohno, Y., Noda, T., & Mikoshiba, K. (2000). Zic2 regulates the kinetics of neurulation. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(4), 1618–1623. https://doi.org/10.1073/pnas.97.4.1618

O'Donnell, N., Zachara, N. E., Hart, G. W., & Marth, J. D. (2004). Ogt-dependent X-chromosome-linked protein glycosylation is a requisite modification in somatic cell function and embryo viability. *Molecular and Cellular Biology*, *24*(4), 1680–90. https://doi.org/10.1128/MCB.24.4.1680-1690.2004

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–21. https://doi.org/10.1101/gr.097857.109

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., … Haussler, D. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, *443*(7108), 167–172. https://doi.org/10.1038/nature05113

Prabhakar, S., Noonan, J. P., Pääbo, S., & Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science (New York, NY)*, *314*(5800), 786. https://doi.org/10.1126/science.1130738

Prabhakar, S., Visel, A., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., … Noonan, J. P. (2008). Human-Specific Gain of Function in a Developmental Enhancer. *Science*, *321*(5894), 1346–1350. https://doi.org/10.1126/science.1159974

Quinn, M. E., Haaning, A., & Ware, S. M. (2012). Preaxial polydactyly caused by Gli3 haploinsufficiency is rescued by Zic3 loss of function in mice. *Human Molecular Genetics*, *21*(8), 1888–1896. https://doi.org/10.1093/hmg/dds002

Rahmann, S., Muller, T., & Vingron, M. (2003). On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol*, *2*, Article7. https://doi.org/10.2202/1544-6115.1032

Reumann, M. K., Strachna, O., Yagerman, S., Torrecilla, D., Kim, J., Doty, S. B., … Mayer-Kuckuk, P. (2011). Loss of transcription factor early growth response gene 1 results in impaired endochondral bone repair. *Bone*, *49*(4), 743–752. https://doi.org/10.1016/j.bone.2011.06.023

Richardson, L., Venkataraman, S., Stevenson, P., Yang, Y., Moss, J., Graham, L., … Armit, C. (2014). EMAGE mouse embryo spatial gene expression database: 2014 update. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1155

Ritter, D. I., Li, Q., Kostka, D., Pollard, K. S., Guo, S., & Chuang, J. H. (2010). The importance of Being Cis: Evolution of Orthologous Fish and Mammalian enhancer activity. *Molecular Biology and Evolution*, *27*(10), 2322–2332. https://doi.org/10.1093/molbev/msq128

Rubin, C.-J. J., Megens, H.-J. J., Barrio, A. M., Maqbool, K., Sayyab, S., Schwochow, D., … Andersson, L. (2012). Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences*, *109*(48), 19529–19536. https://doi.org/10.1073/pnas.1217149109

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., & Furlong, E. E. M. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes and Development*, *21*(4), 436–449. https://doi.org/10.1101/gad.1509007

Sears, K. E., Behringer, R. R., Rasweiler IV, J. J., & Niswander, L. A. (2007). The Evolutionary and Developmental Basis of Parallel Reduction in Mammalian Zeugopod Elements. *The American Naturalist*, *169*(1), 105–117. https://doi.org/10.1086/510259

Shafi, R., Iyer, S. P., Ellies, L. G., O'Donnell, N., Marek, K. W., Chui, D., … Marth, J. D. (2000). The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(11), 5735–9. https://doi.org/10.1073/pnas.100471497

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews. Genetics*, *15*(4), 272–86.

https://doi.org/10.1038/nrg3682

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–50. https://doi.org/10.1101/gr.3715005

Siepel, A., & Haussler, D. (2005). Phylogenetic Hidden Markov Models. *Engineering*, (12), 325–351. https://doi.org/10.1089/1066527041410472

Simmons, N. B., Seymour, K. L., Habersetzer, J., & Gunnell, G. F. (2008). Primitive Early Eocene bat from Wyoming and the evolution of flight and echolocation. *Nature*, *451*(7180), 818–821. https://doi.org/10.1038/nature06549

So, P. L., & Danielian, P. S. (1999). Cloning and expression analysis of a mouse gene related to Drosophila odd-skipped. *Mechanisms of Development*, *84*(1–2), 157–160. https://doi.org/10.1016/S0925-4773(99)00058-1

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., … Birney, E. (2012). Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biology*, *13*(9), R49. https://doi.org/10.1186/gb-2012-13-9-r49

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, *16*(1), 16–23. https://doi.org/10.1093/bioinformatics/16.1.16

Stricker, S., Brieske, N., Haupt, J., & Mundlos, S. (2006). Comparative expression pattern of Odd-skipped related genes Osr1 and Osr2 in chick embryonic development. *Gene Expression Patterns*, *6*(8), 826–834. https://doi.org/10.1016/j.modgep.2006.02.003

Stricker, S., Mathia, S., Haupt, J., Seemann, P., Meier, J., & Mundlos, S. (2012). Odd-skipped related genes regulate differentiation of embryonic limb mesenchyme and bone marrow mesenchymal stromal cells. *Stem Cells and Development*, *21*(4), 623–33. https://doi.org/10.1089/scd.2011.0154

VanderMeer, J. E., & Ahituv, N. (2011). cis-regulatory mutations are a genetic cause of human limb malformations. *Developmental Dynamics*, *240*(5), 920–930. https://doi.org/10.1002/dvdy.22535

Verlinden, L., Kriebitzsch, C., Eelen, G., Van Camp, M., Leyssens, C., Tan, B. K., … Verstuyf, A. (2013). The odd-skipped related genes Osr1 and Osr2 are induced by 1,25-dihydroxyvitamin D3. *Journal of Steroid Biochemistry and Molecular Biology*. https://doi.org/10.1016/j.jsbmb.2012.12.001

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. a, Holt, A., … Pennacchio, L. a.

(2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–8. https://doi.org/10.1038/nature07730

Visel, A., Rubin, E. M., & Pennacchio, L. A. (2009). Genomic views of distant-acting enhancers. *Nature*, *461*(7261), 199–205. https://doi.org/10.1038/nature08451

Wamstad, J. A., Alexander, J. M., Truty, R. M., Shrikumar, A., Li, F., Eilertson, K. E., … Bruneau, B. G. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, *151*(1), 206–20. https://doi.org/10.1016/j.cell.2012.07.035

Wang, Z., Dai, M., Wang, Y., Cooper, K. L., Zhu, T., Dong, D., … Zhang, S. (2014). Unique expression patterns of multiple key genes associated with the evolution of mammalian flight. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1783), 20133133–20133133. https://doi.org/10.1098/rspb.2013.3133

Yang, Y. R., Song, M., Lee, H., Jeon, Y., Choi, E. J., Jang, H. J., … Suh, P. G. (2012). O-GlcNAcase is essential for embryonic development and maintenance of genomic stability. *Aging Cell*, *11*(3), 439–448. https://doi.org/10.1111/j.1474-9726.2012.00801.x

Zhang, G., Cowled, C., Shi, Z., Huang, Z., Bishop-Lilly, K. A., Fang, X., … Wang, J. (2013). Comparative Analysis of Bat Genomes Provides Insight into the Evolution of Flight and Immunity. *Science*, *339*(6118), 456–460. https://doi.org/10.1126/science.1230835

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Jana Friedrich
Author Signature

3/7/2017
Date