Analyzing the Evolutionary Conservation and Tissue-Specificity of Regulatory Clusters

Dilip Thiagarajan

Period 1

June 8, 2015

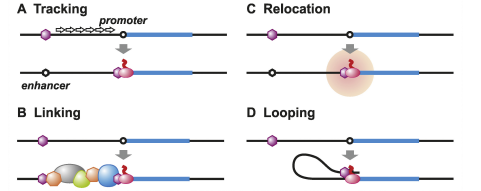
Thomas Jefferson High School, Biotechnology Research Lab

Dr. Andrea Cobb

**Background**

Cellular phenotype, the physical appearance or one such exhibited by the body, is a direct result of multiple regulatory interactions between ribosomes, RNA polymerases, specific DNA coding sequences (exons), and regulatory elements found near exons in the genome. Specifically, regulatory elements are significant in enacting temporal and tissue-specific control on the observed phenotype. In particular, enhancers and silencers have a prominent role in respectively activating and repressing transcription of a gene, as well as being ubiquitous within the non-coding sequences of the genome.

The function of regulatory elements, such as enhancers and silencers, are inherently assumed to be binary, either preventing or allowing transcription of their associated gene. However, this assumption is somewhat incorrect, as delineated in the four existing models of enhancer and silencer function, shown in Figure 1 below.

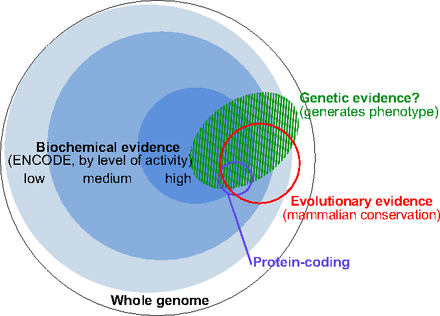


**Figure 1.** Four existing models of enhancer/silencer function in the genome. (Kolovos, Knock, Grosveld et. al.)

In a study by Kvon, Kazmar, Stampfel et. al, enhancers were found to show patterns in movement and activity during development, appearing to regulate the transcription of nearby genes, hinting that these regulatory factors might be organized into groups of interacting elements. Another study showed that enhancers commonly containing short, DNA motifs act as binding sites for proteins necessary for transcription by recruiting proteins that helped to either activate or repress transcription through coordinate activity. (Shlyueva, Stampfel, Stark. 2014.). These two studies are representative of a larger collection of studies, all noting that the non-coding genome possibly interacts within itself and associates outside of its own domain, and that the regulatory elements comprising the non-coding region coordinate activities in a pre-defined manner to produce the correct phenotype.

The ENCODE Consortium (*ENCODE: Encyclopedia of DNA Elements*, 2014), led by the University of California at Santa Cruz, has begun investigating the hidden role of these regulatory elements in their search for all "functional" elements of the human genome. Although they have been frequently criticized for defining a "functional" element very loosely (their general guideline for defining a functional element is shown in the Figure 2), their results have been shown to be significant based on their investigation of regulatory elements (done by running assays of DNA hypersensitivity (the sensitivity of a particular region of DNA to cleavage by DNAse I), DNA methylation, and immunoprecipitation of proteins involved in transcription. (*ENCODE: Encyclopedia of DNA Elements*, 2014)). As a result, they were able to build a database including all of the elements of the human genome on each chromosome.

Although the identification of the elements seems to be comprehensive, a problem with the models that distinguish these elements as shown in Figure 1 insinuate that enhancers and silencers enact regulatory control on their associated genes if and only if they are proximal (within 3000 base pairs of distance) to that specific gene. However, an ENCODE Consortium experiment found that the number of regulatory elements identified significantly outnumbered the amount of genes. Further, these elements were often distal (over 3000 base pairs) from the genes they regulate, refuting the idea that enhancers and silencers must be proximal to a gene in order to regulate it.

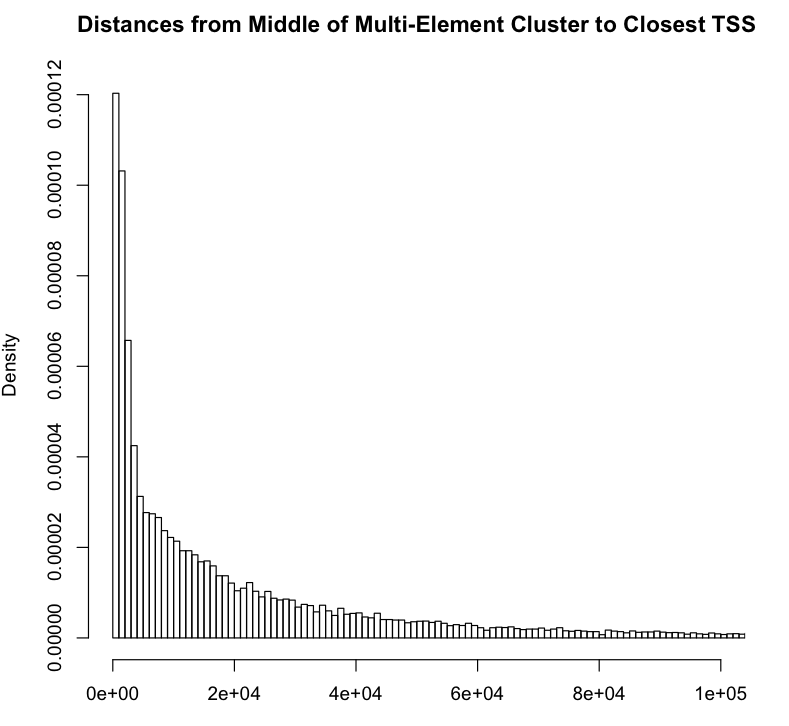


**Figure 2.** The general procedure by which researchers associated with the ENCODE Consortium defined "functional" elements. The general basis was based first on biochemical activity, and then whether or not the gene segment was involved with phenotypic or proteomic expression, or if it was conserved across mammals. (Kellis, Wold, Snyder et. al.)

Research has also been done in an effort to annotate *cis*-regulatory components with respect to the functionality of the associated genes. For example, Ernst and Kellis used a Hidden Markov Model in order to analyze the accessibility of chromatin in human T cells, allowing them to record further enrichments regarding function, motifs, and other experimentally observed characteristics. This research directly provided a greater understanding of the *cis*-regulatory genome and providing useful direction for further study of the non-coding region. Another similar study related chromatin state dynamics (i.e. how the chromatin accessibility in the *cis*-regulatory region and the role of chromatin states in disease. (Ernst, Kheradpour, Mikkelsen et. al.)

Both these studies showed that proximal and distal interactions are present between the *cis*-regulatory genome and exons and have accordingly annotated the *cis*-regulatory genome based on the functionality of flanking genes and protein enrichments. Furthermore, enhancers and silencers collectively occur more frequently and with greater density at points in the genome closer to their target genes.

In my previous research, I determined that these regulatory elements cluster, and that these clusters have genomic significance based on their proximity to their associated genes. In particular, I determined that clusters of regulatory elements are associated with transcriptional starting sites and thus may play a role in transcription, based on their proximity to the nearest transcription starting sites, shown in Figure 3. My previous research and the research I plan on conducting differs from previous studies in that I am addressing the role played by silencers and enhancers, as opposed to published work aimed at studying the role of enhancers only, while also looking at how combinations of the two play a role in genomic function.



Distance (base pairs)

**Figure 3.** Distance from regulatory cluster to most proximal transcription starting site (from my previous research)

However, it is still not clear if the interaction between enhancers and silencers occurs in an efficient, effective, and coordinated manner with respect to the tissue for which their associated gene is expressed in, and for the specific function of that gene. Therefore, with the abundance of data detailing methylation and histone modifications of the products of transcription, such as the Tissue-specific Gene Expression and Regulation (TiGER) database from Johns Hopkins (http://bioinfo.wilmer.jhu.edu/tiger/) and the GREAT database from Stanford (http://bejerano.stanford.edu/great/public/html/), I can surmise the interactions between enhancers and silencers relating more closely to the function of the gene rather than the function of the regulatory elements. In addition, I will investigate the differences during evolutionary turnover between mammals, specifically with the regulatory clusters. The purpose of this study, therefore, is to analyze the significance of clusters of regulatory elements in the function of associated genes based on evolutionary conservation and tissue-specificity.

Going forward, findings from this study would provide for a better understanding of gene regulation in the context of the function of their associated gene. Since the study primarily addresses human gene regulation, it also serves to better our understanding of the manner in which one might modify gene regulation to better combat particular disease conditions.

The goal of this study is to determine whether the regulatory clusters are evolutionarily conserved across mammals, and if differences in conservation do arise between singleton elements and multi-element clusters, which will require that I run an evolutionary conservation analysis. In addition, I will conduct a tissue-specificity analysis (run in tissues similar to those provided in Figure 4) with respect to all the regulatory clusters (conserved and non-conserved amongst mammals) in order to identify significant clusters based on any relation to the transcription factors. Finally, I will conduct a functional analysis to illuminate any possible interactions occurring within the regulatory clusters that were not identified in conservation and tissue-specificity analyses. In order to ensure accuracy, I will run two technical replicates in parallel.

I predict that conservation of clusters will be more significant in mammals more closely related to humans (e.g. chimpanzees will show more conservation than mice with respect to the human genome). Tissue-specificity analyses have shown in previous research prevalence among non-conserved regulatory elements in mammals (Yu, Lin, Zack et. al.), and as a result, I predict the same will occur for non-conserved, regulatory clusters.

**Materials and Methods**

I conducted this study using computational analysis, with the majority of the necessary resources available online. In my previous study, to confirm statistical significance of interactions within regulatory clusters, I retrieved genomic data from the Table Browser from the ENCODE Consortium. I used Python programming and UNIX console commands to build analyses for the genomic data as well as for the data regarding evolutionary conservation and tissue-specificity. I used R to conduct any further statistical analysis. I needed no other software in order to proceed with my analysis, and all databases mentioned from which I retrieved data from are accessible to the general public.

        To begin, I ran an evolutionary conservation analysis, which required data regarding primates and placental mammals (which are the two species most closely related to humans, which is the standard, that have available data) to be taken from the PhastCons database provided by Cornell. This database provides identification of evolutionarily conserved enhancers and silencers across two groups of species being compared, and accordingly provides a conservation scoring for each element that was conserved, based on any annotations associated with the regulatory elements being analyzed. As a result, this provided some notion of how consistent the presence of specific regulatory elements or clusters are across similar species, especially during evolution, when turnover in the genome brings about significant differences in genomic content across different species.

        Given that data has already been taken from the human genome, the species I compared through the evolutionary analysis were the overall genomic data of primates and placental mammals. I used different entries for each species as positive controls to ensure that similarity is more significant in species that are closer to humans. For the negative control, I used genomic data of amphibians.

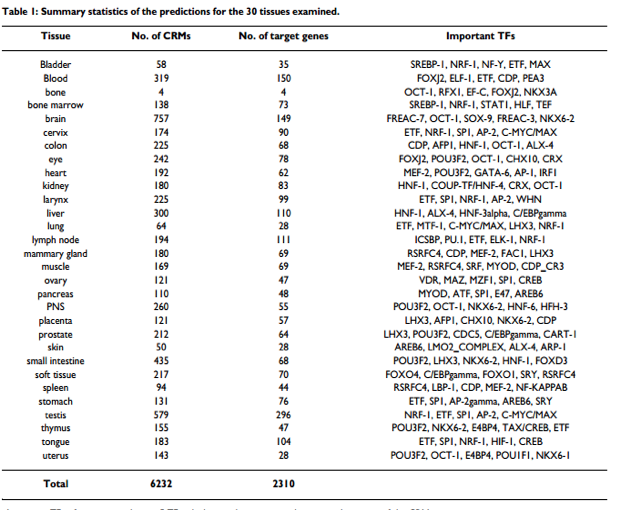
        The Tissue-Specific Gene Expression and Regulation (TiGER) database currently contains tissue-specific expression profiles for approximately 20,000 genes, combinatorial regulation for 7,341 interacting transcription factor (TF) pairs, and 6,232 cis-regulatory modules (CRMs, i.e. regulatory elements) for these tissue-specific genes, and I used these to conduct this analysis. The DNA sequences and annotations available from this database were also retrieved from the UCSC genome browser, so I could compare across multiple entries of genomic data for the tissue-specificity analyses to ensure validity.

        To conduct the actual tissue-specificity analysis, I used the TF view and tissue view in the database to identify all factors that interact with a given TF as well as to identify which genes are preferentially expressed in a given tissue, in order to see any interactions between any TFs and the cis-regulatory modules (CRMs) in the promoter regions of tissue-specific genes. The TFs that I accordingly analyzed were the same as those shown in Figure 4, and from the TF view, I used any annotations frequently associating with those TFs to identify any distal regulatory elements with the same annotation. I then used the tissue view data to try and correlate non-conserved clusters with the clusters of proximal regulatory elements to the associated gene, as in previous research, it was shown that the individual tissue-specific elements tended to be non-conserved elements. (Yu, Lin, Zack et. al.)

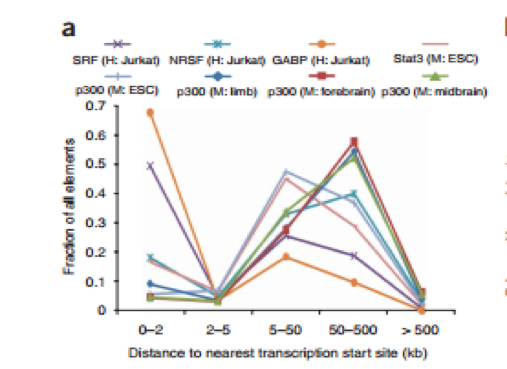
        The next analysis I ran involved examining the distribution of distances, specifically to examine the frequencies of tissue-specific proximal clusters as compared to distal regulatory clusters with annotations frequently associated with important TFs of that tissue. I expected the distance distribution to be similar to that shown in Figure 5, as was found when examining the tissue-specificity of individual enhancers and silencers. (McLean, Bristor, Hiller et. al.)

        In addition, I examined any other annotations not associated with the important TFs that are relevant to the expression of a gene that is directly involved in the function of the tissue. For example, I identified any proteins that increase chromatin accessibility, thereby enhancing expression of the gene, and search for any associations with the regulatory clusters. If any association arose, I recorded the distance of the cluster from the associated gene.

A summary of the procedures I followed is outlined in Figure 6.



**Figure 4.** Important TF factors predicted to be associated with various tissues. (Yu, Lin, Zack et. al.)



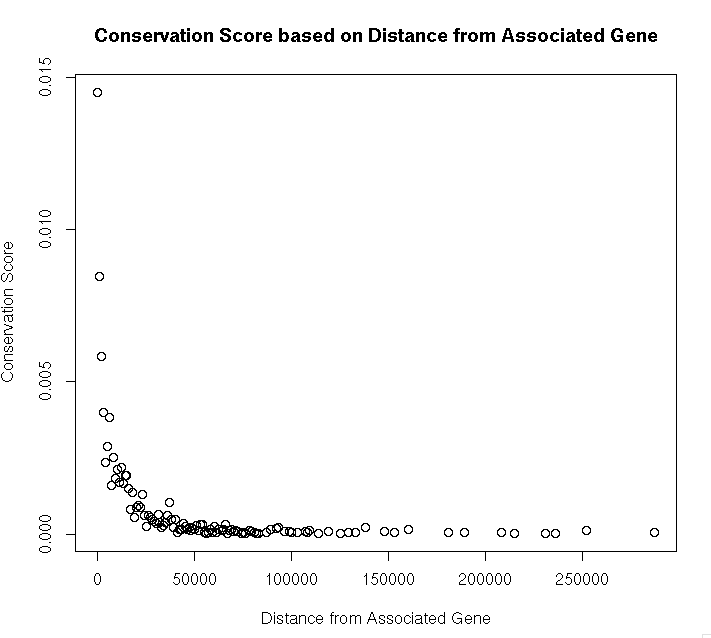
**Figure 5.** Distance distribution of various regulatory elements specific to the tissue. (McLean, Bristor, Hiller et. al.)

        One problem inherent with the limited amount of databases from which I was able to draw data was that there could have been discrepancies between databases detailing different aspects of the same human genome. To compensate for this discrepancy, I used different combinations of genomic data and annotations from the various databases to cross-reference the data before running the separate analyses. In addition, I took genomic data from all the chromosomes rather than drawing random samples of genomic data from a few chromosomes.

**Figure 6.** A summary of the methodology I followed for this study.

**Results**

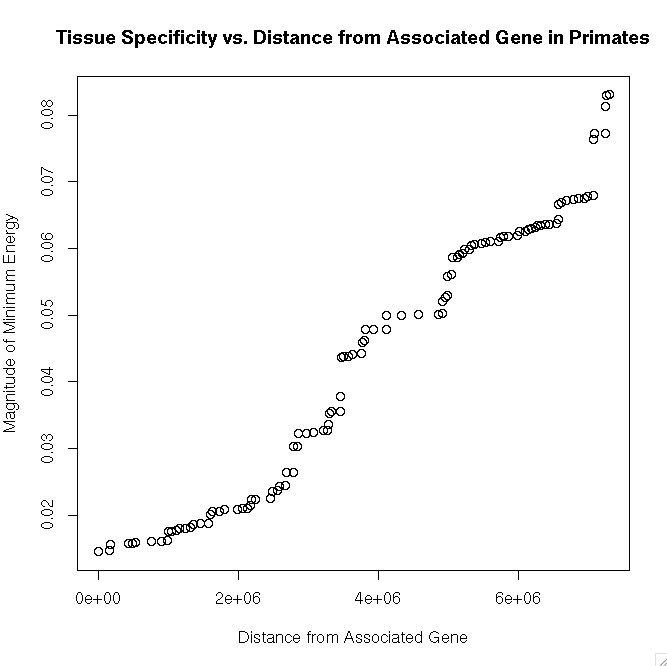
The outcome of the conservation analysis in mammals for clustered regulatory elements followed from previous findings conducted on singleton elements: as the distance from the associated gene increased, the probability of conservation, i.e. the conservation score, would decrease.



**Figure 7.** Conservation score of multi-element clusters of enhancers/silencers vs. distance from the gene associated with the cluster.

Fitting the dataset for conservation scores, as shown in Figure 6, to an inverse linear function gave an R2 value of 0.83, and fitting the dataset to an exponential decay function gave an R2 value of 0.44. However, the best fit occurred when fitting the dataset to an inverse square relation, giving an R2 value of 0.99.

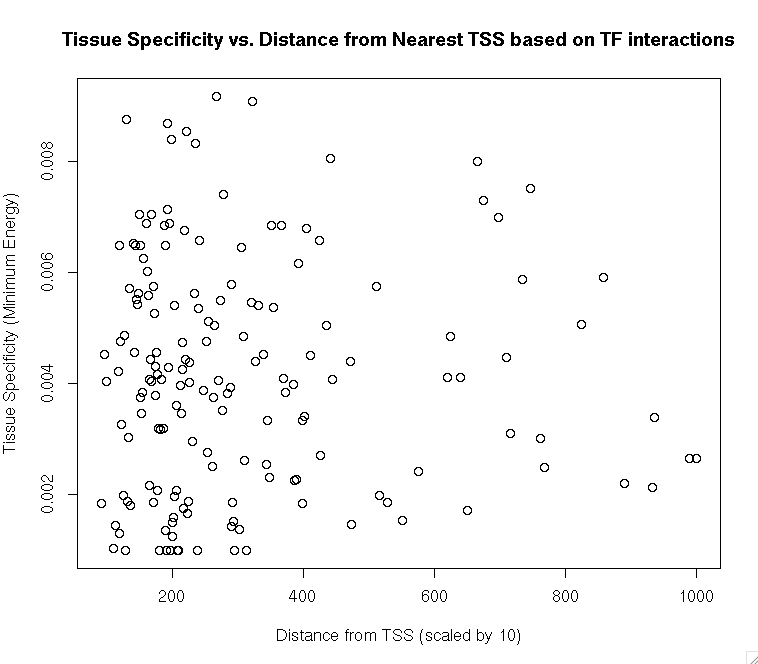
Yu, Lin, Zack et. al. found that in mammals, non-conserved singleton elements (elements with smaller conservation scores) tended to be more tissue-specific, meaning the minimum energy of the association with the gene and the elements was higher than the average minimum energy. This same finding held true for multi-element clusters in primates, the largest subset of mammalian data available from the TiGER database, and the only subset of data that fully encompassed the set of data describing the tissue-specificity of elements for placental mammals.



**Figure 8.**  Tissue specificity (minimum energy) vs. distance from the gene associated with the cluster. Fitting the above dataset to a cubic function shows the strongest correlation.

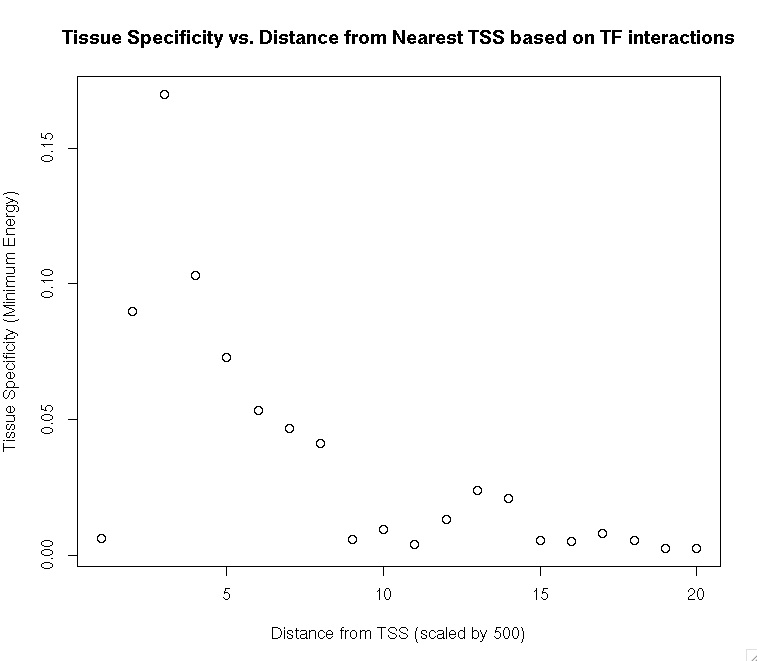
Mapping out the data gave a scatterplot with two clear areas where the tissue specificity did not change over a change in distance from the associated gene, leading immediately to a best fit with a cubic function, with an R2 value of 0.97. The dataset is shown in Figure 7.

The last analysis hinged on the significance of a specific transcription factor (TF) within a given tissue, thereby assessing the functionality of a specific cluster of regulatory elements. Therefore, the distance from the middle of a multi-element, regulatory cluster to the transcription-starting site (TSS) of the associated gene was plotted against the minimum energy of each cluster. The full dataset is shown in Figure 8.



**Figure 9.** Tissue-specificity as measured by TF-affinity vs. distance from the middle of the cluster (scaled by ten base pairs) to the associated transcription start site.

However, given the plethora of data points in the dataset, it was difficult to find a specific correlation or a fitting function. Therefore, I split the dataset into bins of 50 base pairs in order to reduce the clutter and error due to noise of data collection from the database. The minimum energies were averaged within each bin, and the results are shown in Figure 9 below.



**Figure 10.** Tissue-specificity as measured by TF-affinity vs. distance from the middle of the cluster to the associated transcription start site, scaled with bins of 50 base pairs.

This dataset was fit best by a function of the form:

where are all constants, and the R2 being 0.93

**Conclusions**

The purpose of this study was to determine whether regulatory clusters were evolutionarily conserved across mammals, and if differences in conservation do arise between singleton elements and multi-element clusters. This inverse correlation was supported by the results shown in Figure 6, as multi-element clusters do show a decrease in conservation score based on an increase in distance from their associated gene.

One of the major findings that should be investigated in further studies is the exponent of the inverse relation between conservation and distance from the associated gene. The majority of multi-element clusters, as found in the study I conducted preceding this one, contained a pair of elements, and the inverse relation was a square inverse relation. Therefore, it may be fruitful to study the set of clusters with three elements, and see if the inverse relation takes on the form of a cubic inverse, and so on for clusters of larger size.

In addition, I conducted this study to identify any significant multi-element clusters with respect to tissue specificity based purely on tissue-specificity. The tissue-specificity analysis showed a positive, cubic relation, given the appearance of two, distinct troughs. This fit could be related to the function that fit the conservation data as well, and is something worth studying in future research for the subset of clusters with three elements, four elements, and so on.

Finally, I conducted this study to gauge the functional connotation of the significantly tissue-specific clusters. Interestingly, the function that best described the set of data displayed in Figure 4 has no precedent in any research, and is a non-trivial combination of transcendental functions that aren’t normally associated with biological applications. Nevertheless, while the fit and function for the dataset for this analysis, as shown in figure 9, had an R2 value of 0.93, the fit could have been made stronger with more advanced computational resources, and therefore, future studies should work to find a stronger fit for this dataset, using a Taylor series analysis rigorously identify any mathematical relations between the function in this fit and the functions in the tissue-specificity analysis and conservation analysis.

**References**

[ENCODE]. (2003, April). Retrieved September 14, 2014, from Encyclopedia of DNA Elements at UCSC website: http://genome.ucsc.edu/ENCODE/

Provides a "list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active."

*Encode Consortium*. (n.d.). Retrieved November 19, 2014, from https://www.encodeproject.org/

Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *National Biotechnology*, *28*(8), pp. 817-825. http://dx.doi.org/10.1038/nbt.1662

Ernst, J., & Kellis, M. (2013). Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Cold Spring Harbor Laboratory Press*, *23*, pp. 1142-1154. http://dx.doi.org/10.1101/gr.144840.112

Ernst, J., Kheradpour, P., & Mikkelsen, T. S. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*, pp. 43-49. http://dx.doi.org/10.1038/nature09906

Felsenstein, J., & Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution*, *93*(104), 93-104.

*Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project* (ENCODE Consortium, Comp.). (2007, June). Nature. http://dx.doi.org/10.1038/nature05874

Kolovos, P., Knoch, T., Grosveld, F., Cook, P., & Papantonis, A. (n.d.). Enhancers and silencers: An integrated and simple model for their function. *Epigenetics & Chromatin*. http://dx.doi.org/10.1186/1756-8935-5-1

Malin, J., Aniba, M. R., & Hannenhalli, S. (2013, May). *Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers*. http://dx.doi.org/10.1093/nar/gkt374

Margulies, E. H., & Blanchette, M. (2003). Identification and Characterization of Multi-Species Conserved Sequences. Retrieved December 3, 2014, from Genome Research website: http://genome.cshlp.org/content/13/12/2507.abstract

McCord, R. P., Zhou, V. W., & Yuh, T. (2011, December). *Distant cis-regulatory elements in human skeletal muscle differentiation*. http://dx.doi.org/10.1016/j.ygeno.2011.08.003

Mclean, C., & Bristor, D. (2010, May). *GREAT improves functional interpretation of cis-regulatory regions* (M. Hiller, Comp.). http://dx.doi.org/10.1038/nbt.1630

Ong, C.-T., & Corces, V. G. (2014). CTCF: An architectural protein bridging genome topology and function. *Nature Review: Genetics*, *15*, pp. 234-245. http://dx.doi.org/10.1038/nrg3663

Ong, C.-T., & Orces, V. G. (2-11). Enhancer function: New insights into the regulation of tissue-specific gene expression. *Nature Review: Genetics*, *12*(4), pp. 283-293. http://dx.doi.org/10.1038/nrg2957

Sheffield, N., Thurman, R., & Song, L. (2013, March). *Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions*. http://dx.doi.org/10.1101/gr.152140.112

Shlyueva, D., Stampfel, G., & Stark, A. (2014, March 11). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Review: Genetics*, *15*, 272-286. http://dx.doi.org/10.1038/nrg3682

Yu, X., Lin, J., Zack, D. J., & Qian, J. (2006, September). *Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues*. Nucleic Acids Research. http://dx.doi.org/10.1093/nar/gkl595