Identification of "Enhancer and Silencer Jungles" in the Human Genome

Abstract

Genetic control is the source of much phenotypic variation based on the genomic characteristics of any given organism. Such control is usually established through nearby regulatory elements, primarily including enhancers and silencers, and will depend on the properties of and interactions amongst these elements, as well as interactions between these elements and the genomic coding regions. Knowledge of these attributes is required for understanding and modifying the dynamics of gene expression, and so we chose to analyze regulatory elements using data collected by the ENCODE project at UCSC.

Significant regulatory elements of the human genome were filtered based on evolutionary-conserved regions and length, clustered using cutoffs determined from the distribution over their separating distances, and analyzed based on their genomic locations. The role of silencers in cluster formation was examined as well.

We found that regulatory elements, including silencers, often act through spatially clustered groups in the human genome. Additionally, silencers occur closer to clusters of enhancers than would be expected on average, indicating that their role in the formation of these overrepresented clusters is significant. Going forward, the roles of individual enhancers and silencers within these clusters should be examined in an effort to better understand *cis*-regulation.

Introduction

Cellular phenotype, the physical appearance or action exhibited by the body, is a direct result of interaction between ribosomes, RNA polymerases, specific DNA coding sequences (exons), and regulatory elements found near exons in the genome. Regulatory elements, in particular, are noted for being significant in enacting temporal and tissue-specific control on the observed phenotype. Enhancers and silencers are perhaps the most significant of these regulatory elements, due in part to their prominent role in transcription and their ubiquity within the non-coding sequences of the genome.

In a study by Kvon, Kazmar, Stampfel et. al, enhancers were found to "display specific spatial patterns that are highly dynamic during development" and "appear[ed] to regulate their neighbouring genes, suggesting that the *cis*-regulatory [non-coding] factors are organized locally into domains". Another study showed that "enhancer sequences contain short DNA motifs that act as binding sites for sequence-specific transcription," and subsequently, regarding the proteins recruited by enhancers, that "these proteins recruit co-activators and co-repressors such that the combined regulatory cues of all bound factors determine the activity of the enhancer" (Shlyueva, Stampfel, Stark). These two studies are representative of a larger collection of studies, all noting that the non-coding genome is frequently enriched, and that the regulatory elements comprising the non-coding region coordinate activities in a pre-defined manner to produce the correct phenotype.

Several studies have also been done in an effort to annotate *cis*-regulatory components with respect to their functionality. For example, Ernst and Kellis used a Hidden Markov Model in order to "reveal 'chromatin states' in human T cells, based on recurrent and spatially coherent combinations of chromatin marks." This allowed them to show "specific enrichments in functional annotations, sequence motifs, and specific experimentally observed characteristics," directly providing 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 in the *cis*-regulatory region and the role of chromatin states in disease. (Ernst, Kheradpour, Mikkelsen et. al.)

All these studies have shown that long-range 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, it has been observed in passing that enhancers and silencers collectively occur more frequently and with greater density at points in the genome closer to their target genes. However, it is still not clear if, out of all the regulatory elements, enhancers and silencers interact, and if so whether this interaction occurs in an efficient, effective, and coordinated manner.

Therefore, we determined to identify the manner in which these elements would cluster in order, so as to find out how they specifically work together. In particular, we sought to determine if such clusters of regulatory elements are truly associated with transcriptional activity. Unlike in previous studies, we explicitly sought to address the role played by silencers and enhancers, as opposed to focusing only on enhancers. Going forward, findings from this study would provide for a better understanding of gene regulation. 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.

Methods

Given that the basis of this project was to be mostly computational, the majority of the resources used were available for download online. In particular, all of the genomic data used was

retrieved from the Table Browser offered by the University of California at Santa Cruz, and the BEDTools application was used to process the files (all in .bed format) obtained. Further, statistical analysis was required in order to interpret the results of each step and to decide how to proceed throughout the study.

Previous research conducted by other labs usually focused solely on enhancers. As mentioned previously, we sought to additionally address the role of repressors, and in particular silencers, in our examination. We retrieved a collection of human genomic data, including locations of enhancers and silencers, from the Table Browser. (http://genome.ucsc.edu) The human genome sequences used were taken from nine different cell lines: Gm12878, H1hesc, HepG2, Hmec, Hsmm, Huvec, K562, Nhek, and Nhlf. The sequences were contained in files encoded in a .bed format, and accordingly were easy to manipulate using BEDTools. Initially, there were 2,255,761 enhancers (both strong and weak) and 262,485 silencers present across all of the data from these cell lines.

We had sought to uncover informative patterns underlying the distributions of enhancers and silencers, but we were aware that, with so much data, it was quite likely that we would also observe deviations from such patterns. Therefore, in order to reduce noise resulting from transient genetic variation, as well as to consider genetic patterns conserved across multiple species, the enhancers and silencers in our dataset were filtered using evolutionary conserved regions (Loots, Ovcharenko). Each regulatory element was retained for consideration if and only if it shared a common subsequence of at least 50 basepairs (bp) with an evolutionarily-conserved region of the human or mouse genomes. Application of this filter left 1,030,185 enhancers (both strong and weak) and 206,101 silencers retained from the original dataset. At this point, we sought to address a known defect in the original dataset.

The algorithm used by the ENCODE consortium to identify regulatory elements has a tendency to collapse near-continuous elements, regardless of element type, into larger pseudo-elements of length at least 3 kilobasepairs (kbp), and we wished to remove these pseudo-elements from our data set. To this end, a second filter was applied to the remaining elements, such that only enhancers and silencers below 3kbp in length within each cell line were retained. While this filter might have removed some unusually long regulatory elements from our data set, we thought it preferable to use a slightly incomplete dataset than to use a dataset with inaccurate regulatory elements.

After these first rounds of filtering, enhancers and silencers were merged with contiguous elements of the same type (i.e. enhancers were merged with contiguous/overlapping enhancers and silencers were merged with contiguous/overlapping silencers), resulting in a collection of fewer but longer regulatory elements in the two categories. This merging was done within each cell line, after which the genomes of the different cell lines were overlaid and the merging repeated on these overlaid genomes. This process left us with 241,991 enhancers, with mean length 1,033 bp, and 60,793 silencers, with mean length 1,725 bp. We shall hereafter refer to these as the observed regulatory elements.

We wished to determine reasonable measures of proximity for the observed set of enhancers and silencers, and to do this we sought to calculate the distribution of inter-element distances between adjacent elements upon random placement of our observed regulatory elements. This calculation required some effort, as there were many regulatory elements to consider, spread over 23 chromosomes, and the number of possible arrangements within each chromosome varied exponentially with the number of elements to place. Furthermore, an exact calculation of the distribution was computationally and theoretically intractable.

As such, we decided to use an optimized procedure wherein the enhancers were randomly placed one after another on their respective chromosomes, in no particular order, and the distance between adjacent elements calculated. This procedure was then repeated for the silencers, and the pair of procedures was in itself repeated 1,000 times in order to more fully sample the true distribution over inter-element distances. This iterated procedure gave fairly extensive and accurate empirical distributions of inter-element distances for both enhancers and silencers. The lower five percent tails of these two distributions were selected to serve as the cutoff distances for clustering regulatory elements.

Once these cutoffs were obtained, the observed enhancers and silencers were accordingly clustered. In addition, adjacent enhancer-silencer pairs were clustered together if and only if their inter-element distance was exceeded by both cutoff values. The distances between clusters and between elements within the clusters were then calculated to determine which enhancers and silencers might have coordinated regulatory effects on transcription. From here, the role of the silencers was analyzed by first clustering only enhancers together and then finding the distances from silencers to their nearest enhancer cluster. This analysis helped in determining whether or not the enhancers and silencers cooperated in complexes that recruit transcription factors to bind to promoters of genes.

After clustering regulatory elements in this manner, we sought to garner further insight into the regulatory specificity of the clusters by examining the relative positions of these clusters, or jungles, with respect to the exons they regulated. Therefore, the minimum distances from the middle and edge of each jungle to a transcription starting site (TSS) were found. Then, the percentage of jungles spanning multiple gene loci was found, and the relative position of the centers of jungles with respect to the center of the closest gene desert was also found (a gene desert refers to the top 3% of the largest intergenic intervals (Ovcharenko)).

Results

Using the tails of empirically computed distributions over potential inter-element distances, clustering cutoffs of 570 bp and 650 bp were obtained for enhancer and silencer clusters respectively. These cutoff distances are between one half and one third of the mean lengths of enhancers and silencers respectively. These are reasonable ratios, especially since the set of observed regulatory elements by construction contained no pairwise overlapping or contiguous elements of the same type.

Shown in Figure 1 are two graphs comparing the empirical distributions over potential inter-element distances to the actual distributions of inter-element distances. As expected, the empirical distributions, denoted by "enhancer" and "silencer" in the corresponding legends, are skewed to the right as compared to the actual distributions, denoted by "background" in both legends. These graphs show, then, that the actual placement of regulatory elements over their chromosomes allowed for more space between adjacent elements than would be expected from uniformly random placement of regulatory elements.

Using the cutoffs stated above, the elements on the chromosomes were clustered based on the described methodology. Figure 2 shows the general results of the clustering. Specifically, the cluster lengths, defined for each cluster as the distance from the start of the cluster's first element to the end of the cluster's last element, tended to mostly be below 3000 bp in size. Additionally, while there were many singleton clusters (i.e. clusters containing only one element), approximately 40% of the clusters were found to contain more than one element. Of these multiple-element clusters, 80% possessed a pair of elements, namely one enhancer and one silencer.

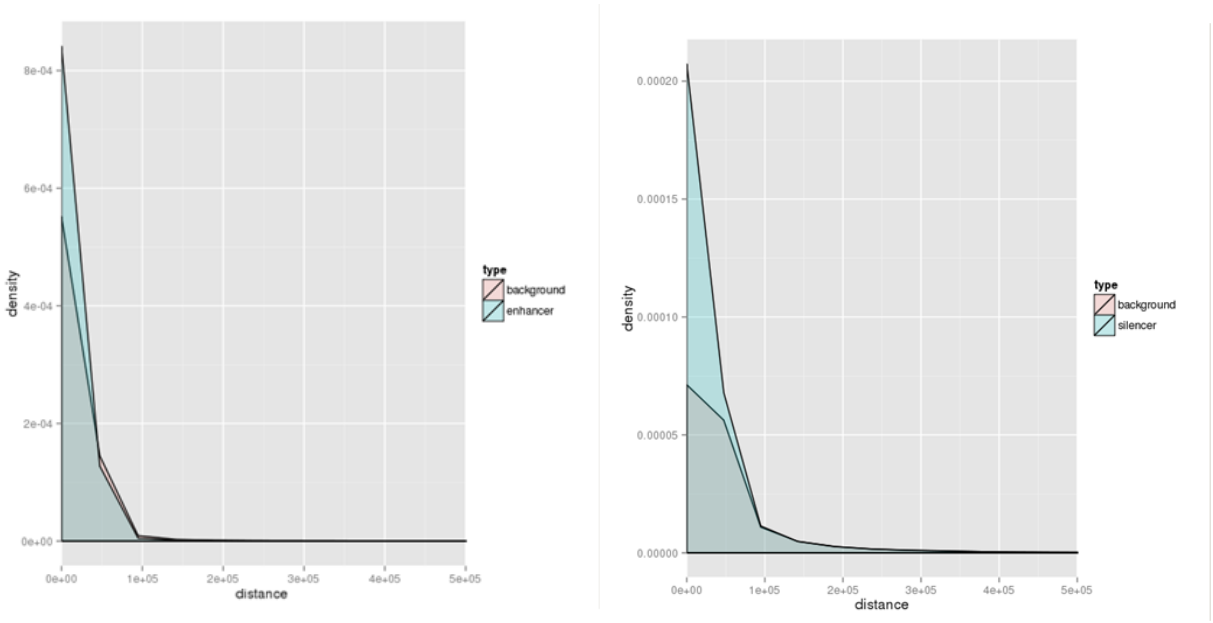

Figure 1. Distribution of inter-element distances between adjacent regulatory elements. The background distribution corresponds to the actual placement of regulatory elements, while the enhancer and silencer distributions were obtained by repeated placement, uniformly at random, of the observed regulatory elements on their chromosomes.

To analyze the clustering, the distances between elements within clusters was found, as well as the distances between clusters. As expected, the maximum distance between elements within a cluster was 649 bp (one base pair below the maximum cutoff distance), and the minimum distance between two clusters was 571 bp (one base pair above the minimum cutoff distance).

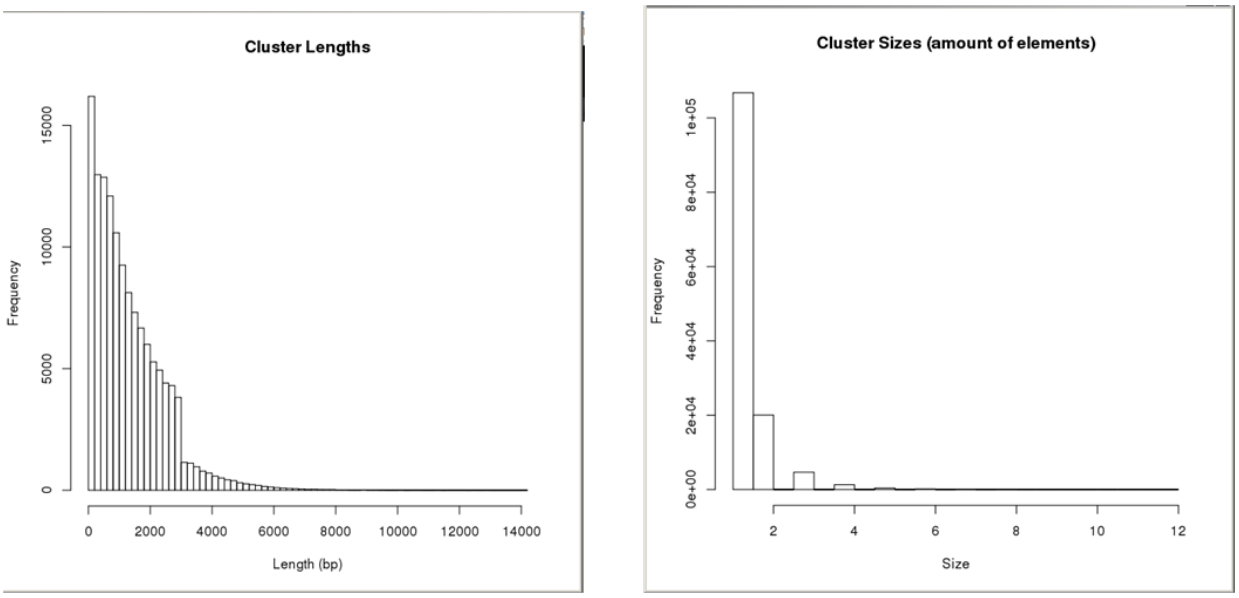

Figure 2. Histograms of cluster lengths and cluster sizes

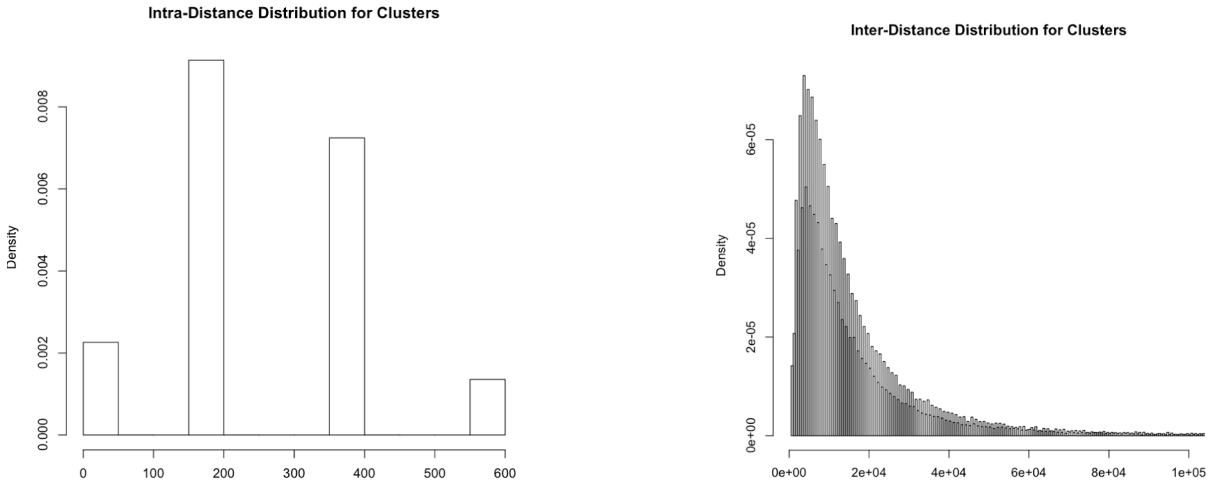

Figure 3. Intra-cluster and Inter-cluster Distributions over Inter-element Distances

The distribution over distances from silencers to the closest, solely enhancer jungles, as seen in Figure 4, is not extremely skewed to the right, but the values tend to be smaller than the average of the distribution. This bias towards smaller values indicates not only that the silencers are generally very proximate to the groups of enhancers, but that enhancer-silencer clusters might be more significant to

use in further analysis, especially when examining tissue-specificity and performing conservation analyses in other vertebrates.

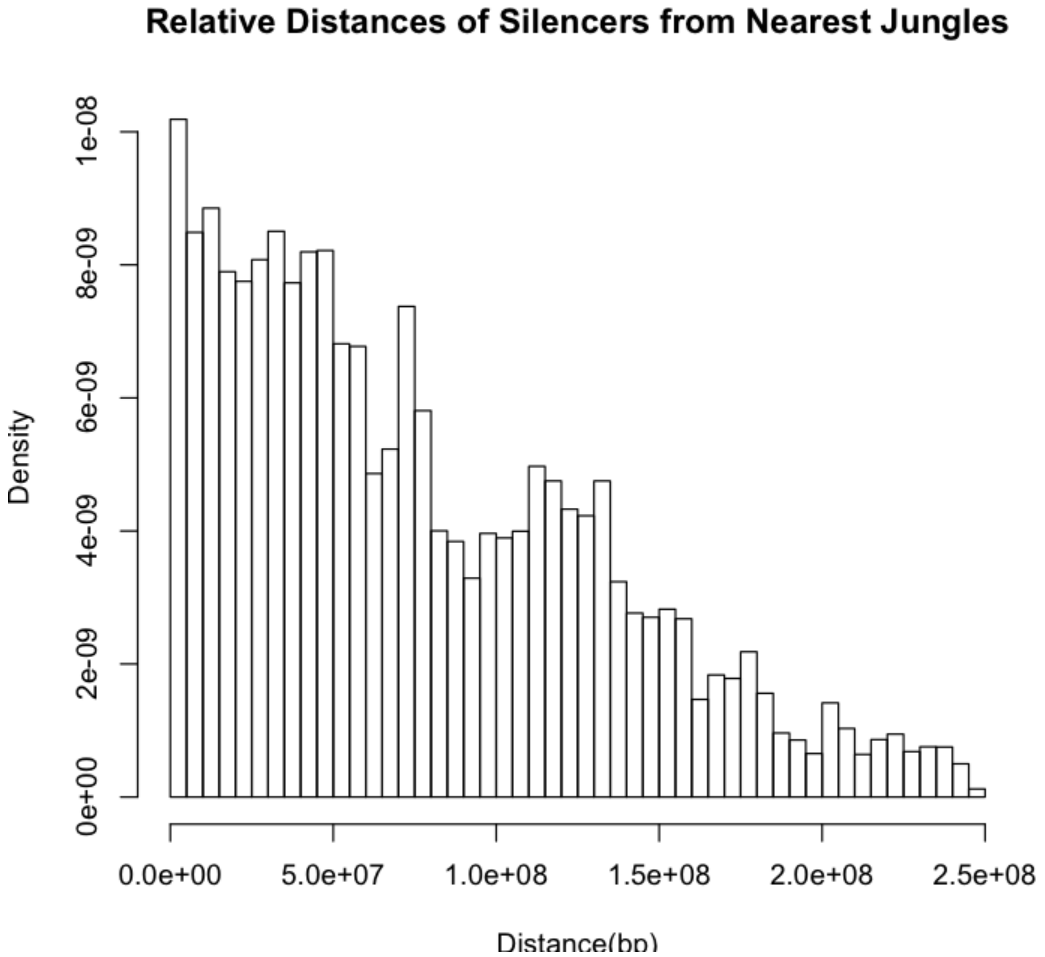

Figure 4. Distances from silencers to nearest enhancer jungles

Figure 5 shows statistics and histograms pertaining to the relative genomic locations of these regulatory jungles.

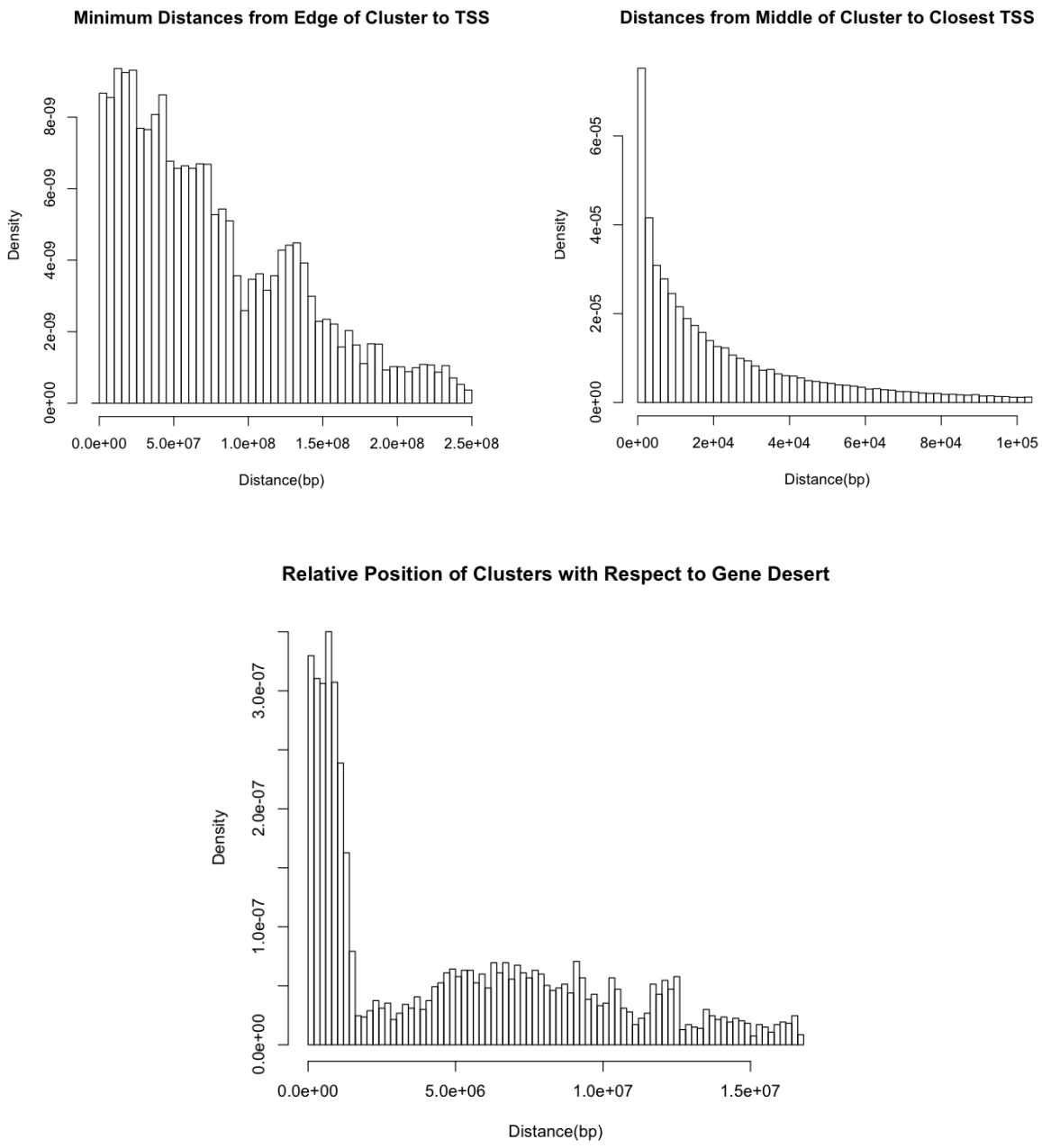

Figure 5. Histograms describing the relative genomic position of the jungles

Discussion

Our initial filtering to remove those regulatory elements that corresponded to evolutionarily transient behavior was expected to have similar effects on both silencers and enhancers. We also expected that most of the elements in the dataset would be filtered out as evolutionary transients during this step. However, we were surprised to find that most of the regulatory elements were retained, indicating that regulatory elements in human and mouse genomes have strong evolutionary pressures associated with them. Furthermore, the fraction of silencers filtered out here was far smaller than the fraction of enhancers filtered out. This served as an initial indicator of the significance of the role the silencers would occupy within the regulatory jungles. It also brought further into contrast the decision made to exclusively analyze enhancers in previous studies. The subsequent filtering and merging of regulatory elements proceeded as anticipated, and did not reveal any telling patterns.

We wished to next determine characteristic cutoff distances for use in grouping regulatory elements into clusters. However, to do so required a theoretical model of distributions over inter-element distances for adjacent elements. At present, no such model has been posited, much less validated, and so we sought to construct an empirical approximation to a presumptive model by repeated placement, uniformly at random, of the observed regulatory elements on their chromosomes, as described in the methodology. Indeed, we attempted to fit both our empirical distribution and the observed distribution with several models, including the Poisson distribution, geometric distribution, and the negative binomial distribution models. None of these candidate models matched either set of data particularly well, especially at larger values of the inter-element distance. We even considered utilizing these candidate models with a restricted domain consisting only of smaller inter-element distances, as the clustering cutoffs would rely in some sense only on the left tail of the chosen distribution. However,

we ultimately decided to use the empirical distribution described in order to avoid an unnecessary and potentially misleading truncation of the dataset.

As seen in Figure 1, the actual distribution tends to have fewer smaller values than the empirical distribution of distances between the elements on the chromosomes. This observation was validated over several random placements of enhancers and silencers onto their respective chromosomes. The empirical distribution superimposed on the actual distribution is shown in Figure 1, and the demonstrated overlap between the two distributions along with the empirical bias towards smaller inter-element distances reaffirmed our choice to use the empirical distribution to obtain the cluster cutoff distances..

As we proceeded to the actual clustering, we observed some subtle but interesting trends. One such trend was the abrupt decrease in frequency of the cluster lengths at the 3 kbp mark. Of course, the second filter applied to our dataset did remove all listed regulatory elements with length greater than 3 kbp, but one might reasonably have expected that the merging of contiguous and overlapping elements within and across cell lines would reintroduce elements of greater length. The fact that this upper bound is maintained, then, was quite interesting, though not definitively significant.

As shown in Figure 2, clustering produced several singletons, or one-element clusters. These singletons, consisting of single regulatory elements with no neighboring regulatory elements within a cutoff distance, made up about 60% of all the clusters obtained. This further supports the hypothesis, first suggested by the paucity of observed regulatory elements of length greater than 3 kbp, that most regulatory elements are sufficiently well-spaced and have length less than 3 kbp.

The singleton clusters were subsequently separated from the multiple-element clusters. Statistics pertaining to inter-element distances for the multiple-element clusters are shown below in Figure 6.

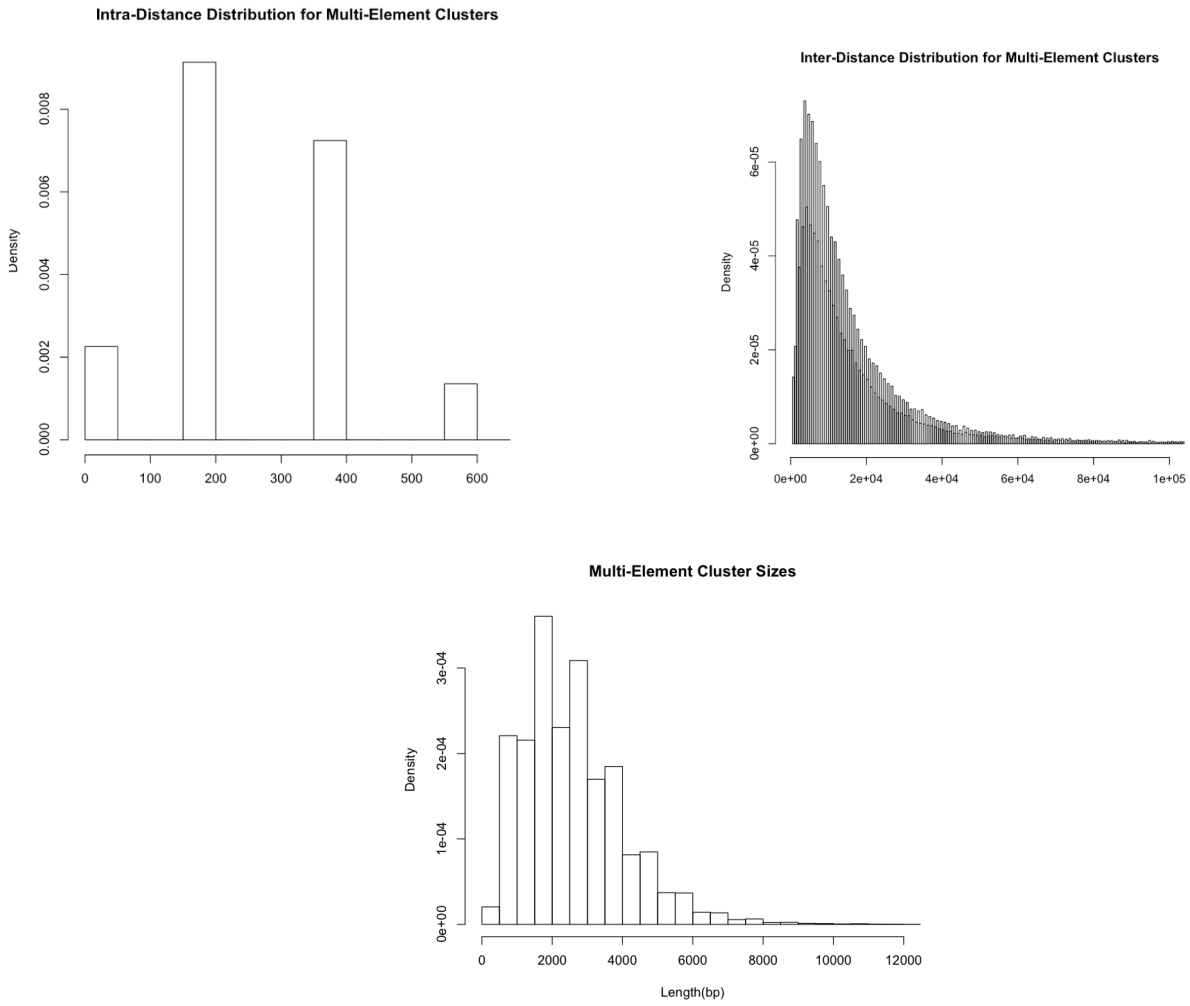

Figure 6. Statistics for multiple-element clusters. The upper left distribution is over inter-element distances within any particular cluster, the upper right distribution is over inter-cluster distances, and the lower distribution is over cluster sizes, all in bp.

Furthermore, analyses pertaining to the locations of multiple-element clusters in relation to regulated exons were performed, and the relevant statistics are shown below in Figure 7.

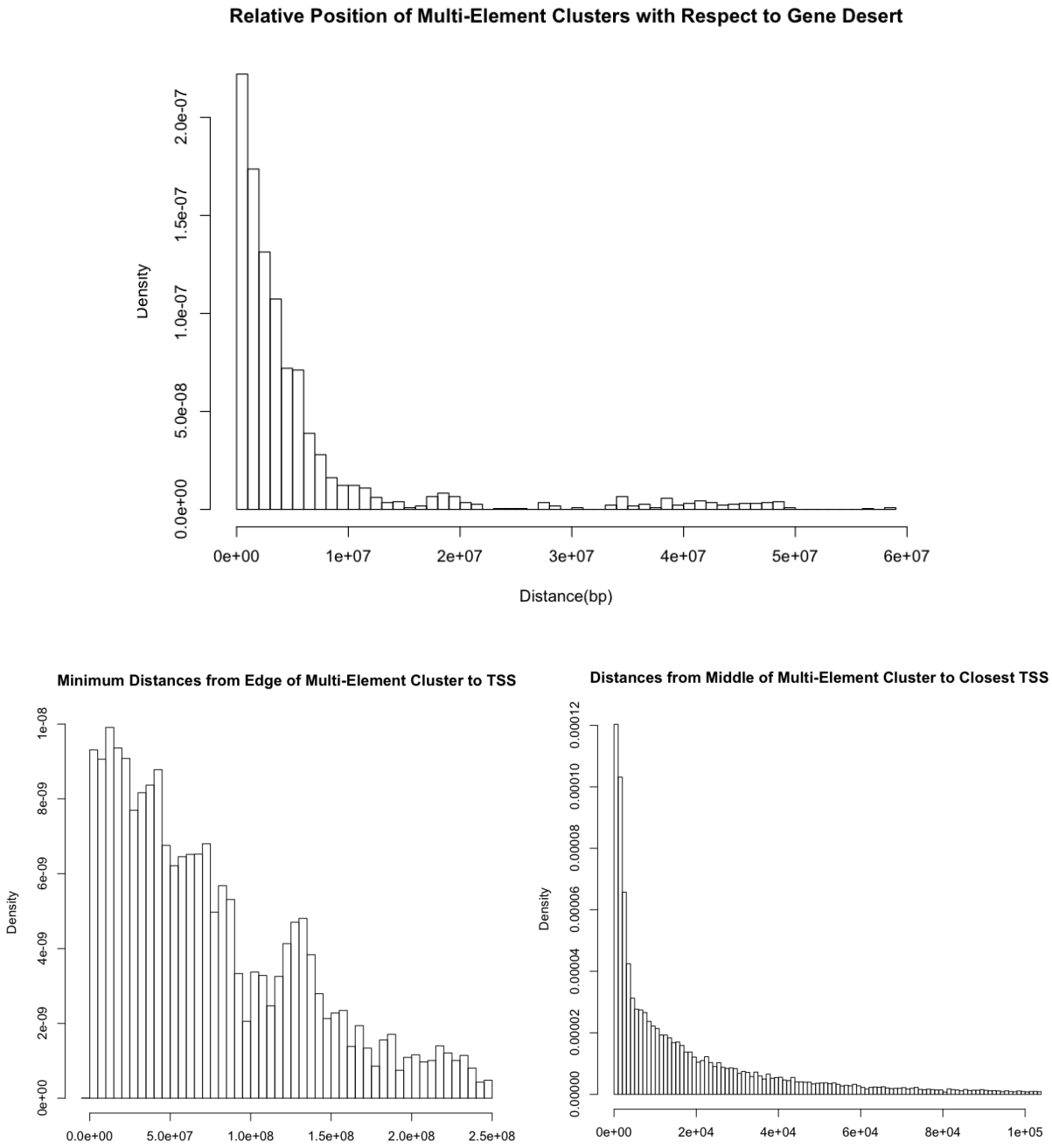

Figure 7. Genomic location analysis of multiple-element clusters. All distances are given in bp.

Finally, the positions of silencers in relation to multiple-element enhancer clusters were examined, and the relevant statistics computed. These are shown below in Figure 8.

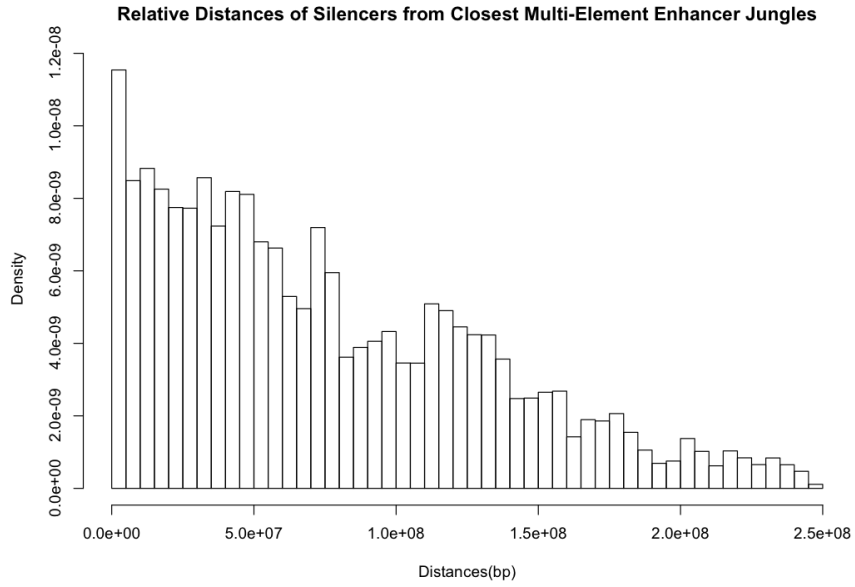

Figure 8. Relative distances of silencers from closest multiple-element enhancer jungles.

Overall, the diagrams do not show much difference between singletons and multiple-element clusters. However, one notable difference lies in the distribution of distances from clusters to gene deserts, which have been shown to "have distinct evolutionary histories and sequence signatures that set them apart from the rest of the genome" (Ovcharenko, Loots). In particular, this distribution is skewed much more to the right for the multiple-element clusters as compared to the singletons. Given that the gene deserts are the top three percent of the largest intergenic intervals, this observation lends itself to a number of possible implications. For example, while multiple-element clusters likely have greater regulatory effect than singletons, this suggests that on average multiple-element clusters might work to regulate fewer genes than do singletons, as proximity to a gene desert corresponds to a smaller number of nearby genes.

Less surprisingly, though no less significantly, distances to the nearest TSS (from both cluster centers and cluster ends) are, on average, smaller for multiple-element clusters as compared to singletons. This strongly suggests the multiple-element clusters are likely to be involved with transcription as one coordinated group, rather than as individual elements, and that this interaction is stronger than would be achieved through independent action of the constituent elements.

Turning our attention for a moment to the distribution of distances between clusters and gene deserts over all clusters, as shown in Figure 5, we note the presence of two peaks, one broad and one narrow, among the higher distance values. An immediate explanation does not present itself for these two peaks, and going forward it would be interesting to examine the compositions of the clusters comprising these two peaks.

In addition to the number of constituent elements, a cluster is also characterized by its size and, more specifically, by the number of genes contained within the cluster boundaries. The percentage of clusters containing multiple genes varied greatly from chromosome to chromosome, and in particular some chromosomes had unexpectedly high percentages of clusters containing multiple genes. The most noteworthy percentages of clusters containing multiple genes are as follows: on chromosome 4, 41%; on chromosome 5, 73%; on chromosome 9, 44%; on chromosome 12, 38%; on chromosome 15, 42%, and on chromosome 19, 36%. This type of data has had no previous analog in other studies, so it would be interesting to look into the biological context regarding these specific chromosomes, and to determine the physiological consequence of such variation between the chromosomes in this attribute.

To begin with, then, we showed that silencers have significant evolutionary context in a species closely related to humans, as per the relatively small fraction of silencers not retained by the filter comparison against human/mice ECRs. We also showed that the silencers are often proximal to

enhancers and are significantly associated with the exons, based on the results of clustering and the positional analysis of the silencers with respect to the TSSs. Therefore, it can be inferred that silencers do indeed have a significant role in working together with enhancers to enact regulatory action. Further studies should aim to include silencers in their analysis while exploring the effects of the *cis*-regulatory genome.

Conclusion

This study shows clearly that silencers have an undeniable role along with the enhancers in the dynamics of the *cis*-regulatory region flanking the coding sequences of the human genome. Additionally, the distances found from the transcription starting sites to various points on the clusters of regulatory elements showed that the clusters play a role in transcription as coordinated entities rather than as individual elements. Therefore, in further studies, it is imperative that silencers be considered along with enhancers when studying how the *cis*-regulatory sector interacts with exons. However, before the silencers are considered in any given study, an evolutionary conservation analysis may be necessary to ensure that the role of the silencers is not just limited to the human genome and the genome of similar species.

In future studies, it would be interesting to investigate and compare the tissue-specificity of both the clusters of elements and the individual elements examined in this study. Additionally, it may prove fruitful to analyze element reshuffling within the clusters during evolution in order to investigate how specific groupings and enrichments of the *cis*-regulatory genome affect transcriptional dynamics. The relation between these dynamics and the structure and composition of regulating clusters might allow for

a much greater understanding of cellular dynamics as a whole, and eventually might lead to an improved ability to modify relevant biological systems, in particular for medical purposes.

References

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., 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

Kvon, E. Z., Kazmar, T., & Stampfel, G. (2014, January). *Genome-scale functional characterization of Drosophila developmental enhancers in vivo*. http://dx.doi.org/10.1038/nature13395

Loots, G., & Ovcharenko, I. (2006, October). *ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes*. Oxford Journals.

Ovcharekno, I., & Loots, G. (2005, January). *Evolution and functional classification of vertebrate gene deserts*. CSH Press.

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