Inferring the Properties of Transcription Factor Regulation

by

Michal R. Grzadkowski
B.Math University of Waterloo (2013)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2016

Author .... Signature redacted .....
Department of Electrical Engineering and Computer Science
October 26, 2015

Certified by. Signature redacted .....
Manolis Kellis
Professor of Computer Science
Thesis Supervisor

Accepted by ..... Signature redacted .....
Professor Leslie A. Kolodziejski
Chair of the Committee on Graduate Students

Inferring the Properties of Transcription Factor Regulation

by

Michal R. Grzadkowski

B.Math University of Waterloo (2013)

Submitted to the Department of Electrical Engineering and Computer Science
on October 26, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

The regulatory targets of transcription factors are costly to directly detect using existing technologies. Many computational models have thus been developed to infer the genes targeted by TFs using gene expression profiles, position weight matrices modeling TF protein binding, histone modifications, and other secondary datasets. We develop a framework for scoring the potential targets of various TFs using models that take the profile of motif hits on the proximity of transcription start sites as input, and describe methods to validate this framework using expression datasets. These models are then extended to include *cis*-regulatory regions inferred from epigenetic data.

Thesis Supervisor: Manolis Kellis
Title: Professor of Computer Science

Acknowledgments

I would like to thank the members of the Kellis Lab, and especially Gerald Quon, Nezar Abdennur, Zhizhuo Zhang, Yongjin Park, and Pouya Kheradpour for their useful comments and feedback during the various stages of preparing this thesis.

Graduate school is time fraught with new anxieties and challenges, but also a once in a lifetime chance to gain new friends and experiences. I would thus like to note the formative role of my friends and peers in shaping my time at MIT so far.

A final thank you is reserved for Manolis Kellis, for granting me the opportunity to participate in the research in his lab and the many exciting opportunities it has brought!

Contents

List of Figures

Chapter 1

Introduction

Recent years have brought impressive advances in our ability to quantify the constituent elements of gene regulation. Expression is now measured using next-generation techniques such as RNA-seq in lieu of micro-arrays, allowing for more reliable snapshots of which genomic sequences are being transcribed in a given tissue context. Chromatin immunoprecipitation (ChIP) assays have allowed us to detect the histone modifications that change the structure of chromatin, thus activating or repressing certain genes via *cis*-regulatory elements (CREs) such as promoters, enhancers, insulators, and silencers. ChIP assays have also led to profiling of sites where transcription factors (TFs), the most important *trans*-acting factors, bind to the genome. Despite these achievements, the core question of gene regulation remains largely unanswered: when and where do *trans*-regulatory elements interact with *cis*-regulatory elements to modulate the transcription of a target gene?

It is well understood that the nuances of these interactions drive many fundamental biological processes, including tissue differentiation (Shibata et al., 2015), evolution (Spitz and Duboule, 2008), and disease development (Lee and Young, 2013). However, attempts to describe the machinery of gene regulation have so far sacrificed depth for breadth, or vice versa. Due to the present-day impossibility of directly assaying the thousands of interactions between TFs and CREs that occur in a tissue

type under any given condition, experimental studies in this direction tend to focus on a single family of TFs and the CREs associated with a small set of genes (Guo et al., 2013; Kim et al., 2014; Patel et al., 2014). On the other hand, computational approaches which seek to model gene regulatory circuits in their entirety are forced to make crude approximations about the properties of *trans*- and *cis*-regulation and the interaction between the two in lieu of *in vivo* obervations.

For example, Neph et al. (2012) built regulatory networks based on the simplistic assumption that a single TF binding site coinciding with a tissue-specific DNaseI footprint within 5kb upstream or downstream of a gene's transcription start site (TSS) is a sufficient and neccessary condition for a regulatory link between the TF and the gene. A similar approach was undertaken by Glass et al. (2015), who defined the promoter region where binding sites are able to mediate target gene expression as being demarcated by boundaries located 750 base pairs upstream and 250 base pairs downstream of TSSs. These models thus fail to take into account the very real possibility that the binding profiles of transcription factors in the proximity of genes that they regulate may vary from factor to factor and from tissue to tissue.

In this thesis we introduce a model that not only seeks to reproduces the interplay between *trans*-regulatory elements and *cis*-regulatory elements across multiple transcription factors, but also characterizes the position of binding sites around transcription start sites of target genes unique to each TF. This is made possible by the availability of tissue-specific *cis*-regulatory region calls, as well as a curated library of TF binding motifs and their "shuffled" counterparts that allow us to correct binding profiles for background nucleotide bias. Furthermore, unlike many previously described models of regulation, we directly interrogate the biological relevance of our findings by developing metrics that test the concordance of the regulatory links our model reports and tissue-specific transcript expression data produced by next-generation sequencing.

The remainder of this thesis is organized as follows. In Chapter 2, we demonstrate

the shortcomings of traditional methods of identifying transcription factor regulatory potential using motif hits, and introduce new computational approaches to address them, thus allowing us to construct robust binding site profiles by fitting motif hits against expression data. In Chapter 3, we describe how regulatory region calls can be used to improve these profiles' ability to model *trans*-regulation, and consider how what proportion of regulatory activity is driven by global properties of TF proteins, tissue-specific epigenetic modifications, and the interaction between the two.

Chapter 2

Identifying Motif Hit Profiles Associated with *Trans*-Regulation

The process by which transcription factors control the expression of the target genes they regulate is well understood in general but difficult to describe in detail. Each TF produces a protein that has an affinity for a particular DNA sequence; when this binding overlaps with *cis*-regulatory elements (CREs) such as promoters and enhancers associated with a gene, its expression is either decreased (repressed) or increased (activated). However, our knowledge of where binding of *trans*-acting factors to CREs occurs and which CREs are associated with each gene remains limited. Although TF binding can be directly observed through methods such as ChIP-seq (Landt and Marinov, 2012), and regulatory regions can be linked to target genes through methods such as Hi-C (Dekker et al., 2013), both of these assays remain too costly to perform on a genome-wide scale, the former especially for a large number of factors.

These obstables have inspired the development of a variety of computational models of *trans*-regulation that seek to infer binding activity through datasets that are more readily available. An early example of such a model was presented in Nachman et al. (2004), in which the observed expression of a set of genes over a period of time was modelled using regulator-target interactions. The method for doing so involved

using a set of kinetic equations that describe how the rate at which a gene is transcribed is affected by the concentration of an active regulator protein. This model was extended to multiple regulators targeting many genes simultaenously and then optimized over a dataset profiling expression over several stages of the yeast cell cycle. A similar approach was later used to find the *cis*-regulatory modules associated with subsets of genes that shared similar profiles of transcription in yeast (Nachman and Regev, 2009).

However, modeling *trans*-regulation in more complex organisms such as humans presents a set of challenges not present when studying the same mechanisms in yeast. Regulatory circuits are often specific to certain tissue types, driving their differentiation in fetal tissue and controlling the processes unique to each tissue in mature stages. Obtaining expression at multiple time points from most adult human tissues is usually impossible due to ethical concerns, and so we must rely on static snapshots of expression. Nevertheless, statistical models have been developed to describe the relationship between expression profiles and transcription factor activity as measured by ChIP-seq peaks in close proximity to transcription start sites (TSSs) (Cheng et al., 2012). An alternative approach presented in Neph et al. (2012) makes do without ChIP-seq data by relying upon TF motifs and DNaseI footprints instead, using the structural properties of the regulatory networks thus constructed to validate their approach in lieu of concordance with expression.

We introduce a novel approach to inferring properties of *trans*- and *cis*-regulation using static expression datasets. A key feature of our model is that it leverages motif hit data that can be obtained computationally instead of relying on direct observation of TF binding using ChIP-seq reads. These hits are generated using a curated set of motifs that include "shuffled" versions of each motif, allowing us to take the background nucleotide bias into account to improve the ability of these hits to predict TF occupancy. The activation or repression of expression levels in genes identified as targets of a given TF is used to validate our model in lieu of observing

expression over time to infer regulatory relationships. Our model is further enhanced using a combination of tissue-specific regulatory region calls as well as gene structure data, which will be introduced in the next chapter.

2.1 Expression Metrics for Validating Regulatory Relationships

Central to our method of constructing a model of *trans*-regulation is the process of validating candidate models against an expression dataset. We thus choose the tissue types we study based on the availability of high-quality expression data from a large number of samples. As described in Appendix A.1, an expression datasets consisting of 320 lung tissue samples was collected and processed as part of the GTEx project. This expression data is measured from transcripts, not genes, which allows for greater accuracy in identifying genomic locations proximal to the regions that code for the expression profiles on hand.

For the time being we will focus on the lung tissue dataset for demonstrative purposes, before moving on to the other two datasets. This lung dataset consists of expression level measurements taken of 57430 transcripts from the 320 samples. The expression profiles of these transcripts follow a characteristic pattern, where expression variation rises in tandem with mean transcript expression up to a certain point, and then drops for highly-expressed transcripts associated with housekeeping genes (Figure 2-1). Almost all of the 20306 transcripts that are expressed at meaningful levels correspond to protein-coding genes, of which transcription factors are a subset. Since these genes are the ones most likely to be targets of *trans*-regulation, we only consider these transcripts in further analysis. We further cull this set of transcripts to the 19273 located on one of the twenty-two non-sex chromosomes, in order to avoid confounding our findings with gender-specific expression profiles.

Before we construct a model of regulation, we must first decide how to validate

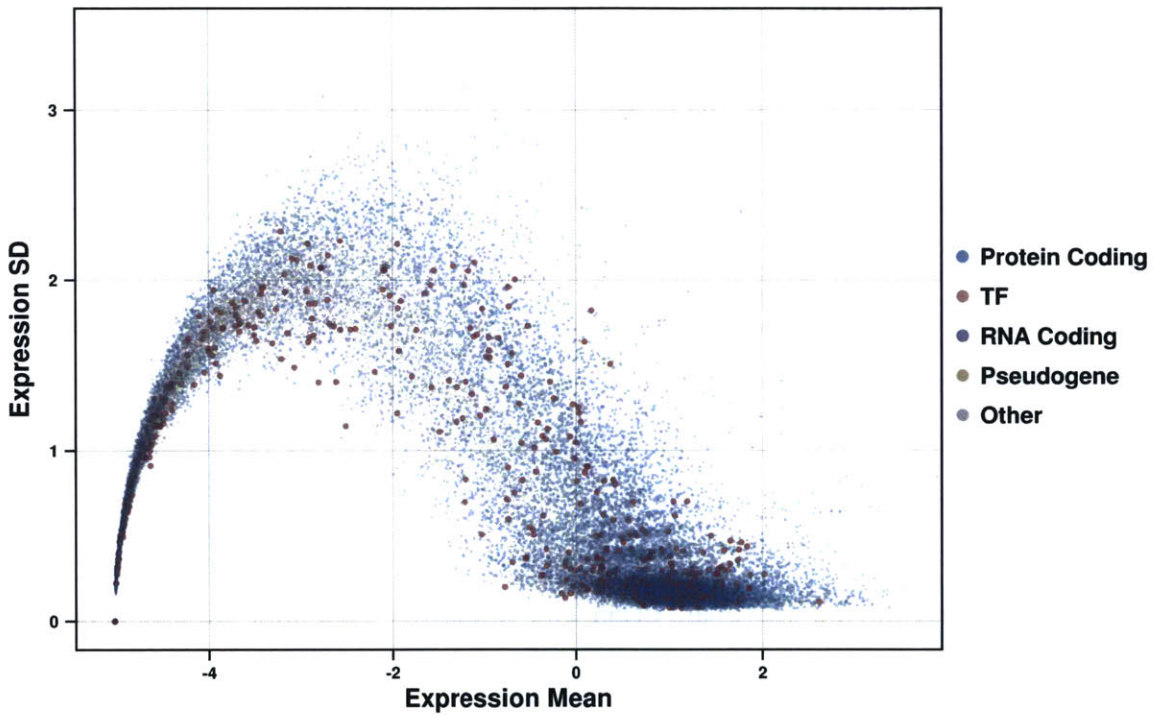

Figure 2-1: Expression profiles of the 57430 transcripts included in the subset of the GTEx expression dataset drawn from lung tissue. The maximally expressed transcript in lung tissue was used for each RNA product.

such a model using a static expression dataset. If a transcription factor does in fact regulate a target gene in the same context in which expression is observed, we should expect the expression of the target gene to be either decreased, if the TF is a repressor, or increased, if the TF is an activator. Assuming that each TF is either a repressor or an activator in a given context, the targets of each TF should thus exhibit consistently lower or higher expression levels relative to other genes. This will be henceforth referred to as a first-order expression effect of *trans*-regulation.

We can further expect that if a transcription factor is highly expressed in a given context or in a given sample, its regulatory effect on the genes it targets will be stronger than in contexts where it is expressed at a lower level, as more of the proteins coded for by the TF will be available to bind to genomic targets. This second-order effect can be measured using the Spearman coexpression rho between TFs and target genes. In particular, genes identified as targets of an activating transcription factor by a model of *trans*-regulation can be expected to exhibit higher coexpression with the TF than other genes, and targets of repressive TFs can be expected to have lower coexpression.

The way in which first- and second-order expression effects are measured depends on how a given regulatory model quantifies the regulatory strength of a transcription factor acting upon a target gene. When regulation is presented as a binary effect, i.e. a link between a TF and a gene is either present or absent, we use the difference in mean expression between the set of genes where the link is present and the set where the link is present to measure first-order effects. This is equivalent to $\log_1$ 0-fold change in expression as our expression dataset in log-normalized.

Likewise, second-order effects for binary models will be measured using the difference in mean coexpression with the TF in question between the set of genes that have been marked as targets and those that haven't. Because genes that are expressed at low levels do not exhibit meaningful differences in expression between samples, their observed coexpression with potential regulators is likely driven by noise rather than

actual interaction *in vivo*, and thus we will exclude these genes when measuring second order effects.

The predicted strength of an interaction between a regulator and a target as returned by a model of interaction can also be continuous. In this case we will use a variety of methods to measure the performance of the given model, which will be described in greater detail in subsequent chapters.

Many processes influence the expression of a gene in a given context other than the binding of *trans*-acting factors, including post-translational modifications, endocrine signalling, and intracellular signals. It is therefore unreasonable to expect the first- or second- order effects associated with any single transcription factor to be particularly large, especially when multiple TFs are involved in the regulation of any given gene as is often the case. We thus focus on predicting TF-gene links that produce better expression effects than various background models, rather than expecting to explain the bulk of the variation inherent in gene expression.

2.2 Using Motif Hits to Identify Targets of *Trans*-Regulation

In order to build a network describing the regulation of target genes by transcription factors, we have to determine which genes are regulated by each TF in a given context. This can be done *in vivo* through methods such as ChIP-seq; however, at present these techniques are sufficiently costly and time-consuming that only a handful of TFs are usually profiled in a given experiment. Several computational approaches for detecting transcription factor binding have thus been proposed, the most prominent of which rely on the use of position weight matrices (PWMs).

A PWM for a given transcription factor, also referred to as a motif, is a model that describes the probability of a sequence of nucleotides of fixed length occuring at a position in the genome where the transcription factor binds (Figure 2-2). Given a

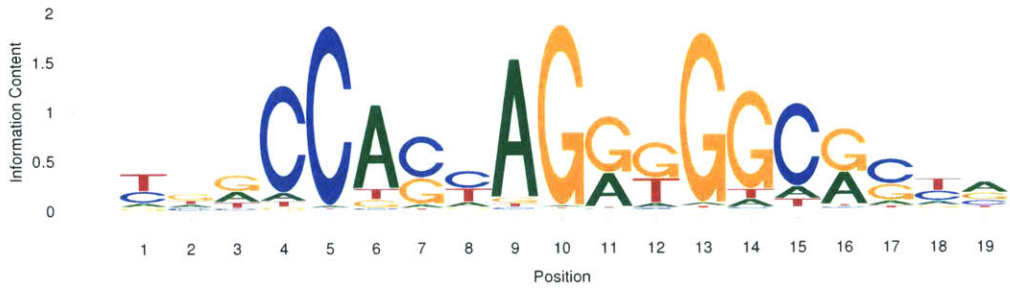

Figure 2-2: The canonical motif for the CTCF zinc finger protein. The height of the column at each position represents the total information content at each position in the motif, with the height of each letter in a column proportional to its probability of appearing at the corresponding position in a sequence occupied by the protein.

motif, we can thus scan any portion of the genome, and assign a score to each genomic position corresponding to the concordance of the sequence at the position with the sequence probabilities encoded by the motif. For the purposes of this study, we will be using the set of motifs collected in (Kheradpour and Kellis, 2014), which is composed of "known" motifs presented in the literature and motifs that were discovered by analyzing fragments of the genome identified as being occupied by transcription factor motifs in recently released ChIP-seq data.

For each motif, we scan the entire genome and identify locations, called motif hits, where the match between the genomic sequence and the motif satisfies a particular p-value threshold based on how often we can expect the motif to match the genome based on chance alone. Each hit is assigned a score between 0 (minimal match necessary to satisfy the p-value threshold) and 1 (best match possible given the motif). We call the set of hits produced by some motif occuring in a particular area of the genome a motif hit profile, fundamentally characterized by the locations and hit scores associated with each hit. Further information on the motifs and methods used to scan the genome for motif scores is available in Appendix A.2.

Given a set of motifs associated with a transcription factor, we must decide how to use the profile of hits for each motif to determine which genes are regulated by the TF. The typical approach to this problem consists of three steps. The first step is to

define a p-value threshold for motif scores to obtain motif hits, as described above. A proximity window around the TSS of target genes corresponding to promoter activity is then chosen, and hits not occuring within these windows are filtered out. A transcription factor is thus linked to a target gene if one ore more of its motifs has at least one hit in the proximity of the gene's TSS. We refer to this approach as the "presence" model, as it defines the motif hit profile in the vincinity of a gene's transcription start site to be proof of *trans*-regulation by the associated TF if at least one hit satisfying the score threshold and window filter is present in the profile.

For example, in one of the first attempts to provide a comprehensive description of tissue-specific *trans*-regulation, (Neph et al., 2012) filtered motif scores using a p-value of 1×10^{-5} and then filtered motif hits using a proximity window with edges 5kb upstream and downstream of the TSS. Tissue-specific regulatory links were generated by further filtering motif hits according to overlap with DNase1 footprints indicating protein occupancy for each tissue. To produce a network model for angiogenesis in ovarian cancer, (Glass et al., 2015) used the same cutoff for motif scores (which were calculated, however, using a different application), but called hits as proximal if they fell within 750 base pairs upstream and 250 base pairs downstream of the TSS. We refer to these as "presence" models, because the presence of single hit satisfying both score and proximity thresholds in the neighbourhood of a gene's TSS is sufficient to link the TF to the gene.

The CENTIPEDE algorithm for predicting transcription factor binding introduces a twist on this approach: motif scores and TSS proximity, in conjunction with evolutionary sequence conservation scores, are used as continuous priors which are fit to binding data as represented by ChIP-seq reads (Pique-Regi et al., 2011). A weakness of the presence paradigm is thus exposed: by transforming motif scores and locations relative to the TSS into binary data by applying hard thresholds, a great deal of information is lost. In both of the studies described above, the regulation of a gene by a regulatory factor is presented as a true/false proposition, with no middle ground,

even though it is almost certainly the case that *trans*-regulation is a wide spectrum, with regulatory relationships ranging from very weak to very strong across different contexts.

Furthermore, the presence of a single motif hit satisfying some set of thresholds hardly seems as sufficient proof of a regulatory relationship. As suggested by the results of the CENTIPEDE study, motif hits closer to a TSS should be given greater weight. The number of motif hits should also be taken into account - but is a gene with a large number of weakly scoring motif hits more tightly regulated by the corresponding TF than a gene with a small number of hits with high scores? Other sources of information not considered in the above studies can also be exploited, such as the strand orientation of each motif hit relative to the target gene, whether motif hits are predominantly upstream or downstream of the TSS, and the overlap of motif hits with *cis*-regulatory elements not necessarily directly adjacent TSSs as identified by analyzing histone marks.

However, before we can construct a more comprehensive model of *trans*-regulation, we must first decide on a valid method of validating such a model for biological relevance. The three studies we have discussed all vary in their approach to validation: Neph et al. analyzed the structural properties of the networks they construct, Glass et al. studied the differential expression of genes identified as being targeted by TFs, and Pique-Regi et al. used a combination of GO term enrichment of target genes and differential expression. We have already introduced a way of measuring regulatory effect through changes in expression and coexpression patters in genes identified as potential targets of transcription factors. We now apply this approach to interrogate the usefulness of the presence model for scoring motif hit profiles for evidence of regulatory activity.

2.3 Validating Motif Presence Models in Lung Tissue

We first examine the effect that the presence of the homocysteine-responsive endoplasmic reticulum-resident ubiquitin-like domain member 1 protein (HERPUD1) has on the expression profiles of putative targets. We choose this TF because it has high mean expression in our lung dataset, ranking second among all TFs, and it is associated with an information-rich motif (Figure 2-3) whose motif hits are relatively sparse across the genome, thus facilitating analysis.

To observe the first-order expression effects of these three motifs, we first use a simplified variant of the presence model presented by Neph et al.. A gene is identified as a target of HERPUD1 if at least one of the motif's hits passing a $p = 1 \times 10^{-5}$ threshold is present within 5kb of the gene's TSS in either direction. To validate the biological relevance of this approach, we measure the difference in mean expression levels between genes that are identified as being targeted by HERPUD1 using these genes and those that aren't. We also repeated the same analysis using a window of 1kb about TSSs in either direction to test the sensitivity of expression effects to distance of motif hits from TSSs.

As shown Figure 2-4, there is no statistically significant difference in expression between the 722 genes that have the HERPUD1 motif present within 5kb of their transcription start site and the 18551 that do not ($p = 0.996$, Wilcoxon rank-sum test). When we use the tighter window of 1kb about the TSS for filtering motif hits, we do observe a much greater absolute decrease in expression in the 199 genes thus identified as being targeted by HERPUD1; however, this decrease also lacks statistical significance ($p = 0.691$). The mere presence of a motif hit in the vincinity of gene TSSs is thus insufficient to produce a noticeable regulatory effect in the case of HERPUD1.

Before proceeding further, we note that our use of first-order expression effects

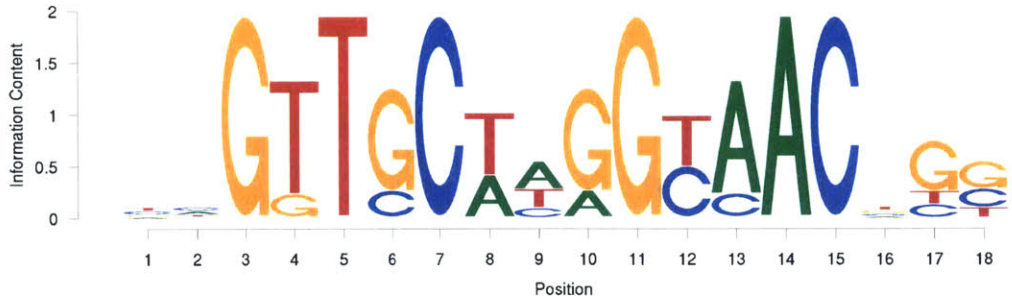

Figure 2-3: The motif associated with HERPUD1.

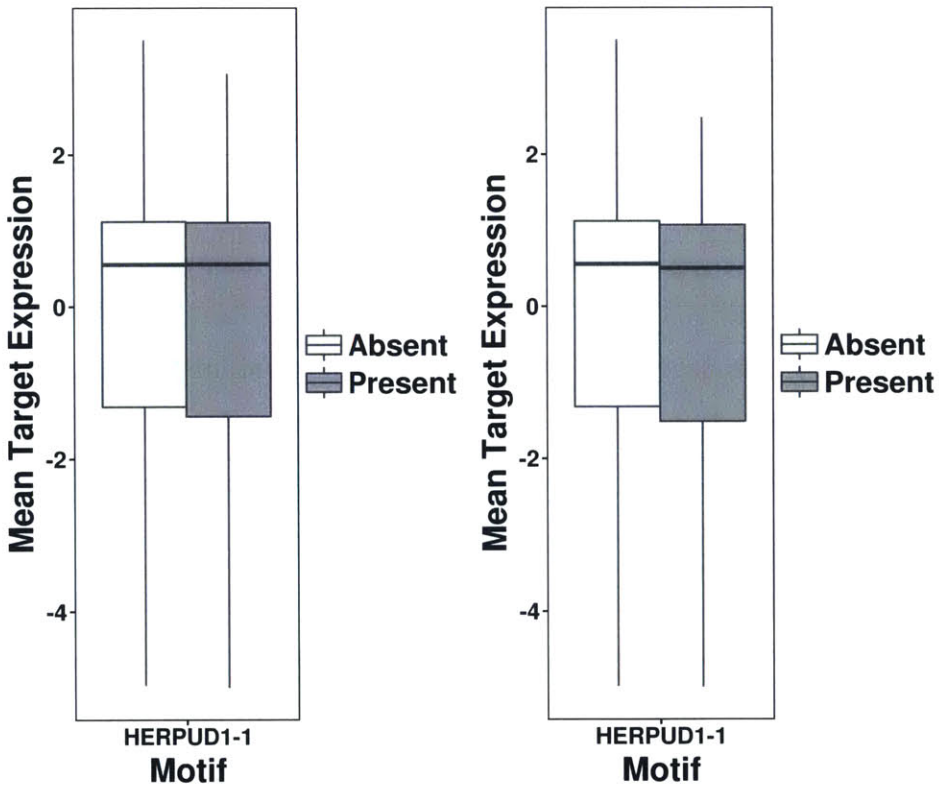

Figure 2-4: First-order expression effects for the HERPUD1 motif using a $p = 1 \times 10^{-5}$ threshold for motif hits and a 5kb window around each transcription start site (left) and a 1kb window (right) to measure *trans*-regulatory effect.

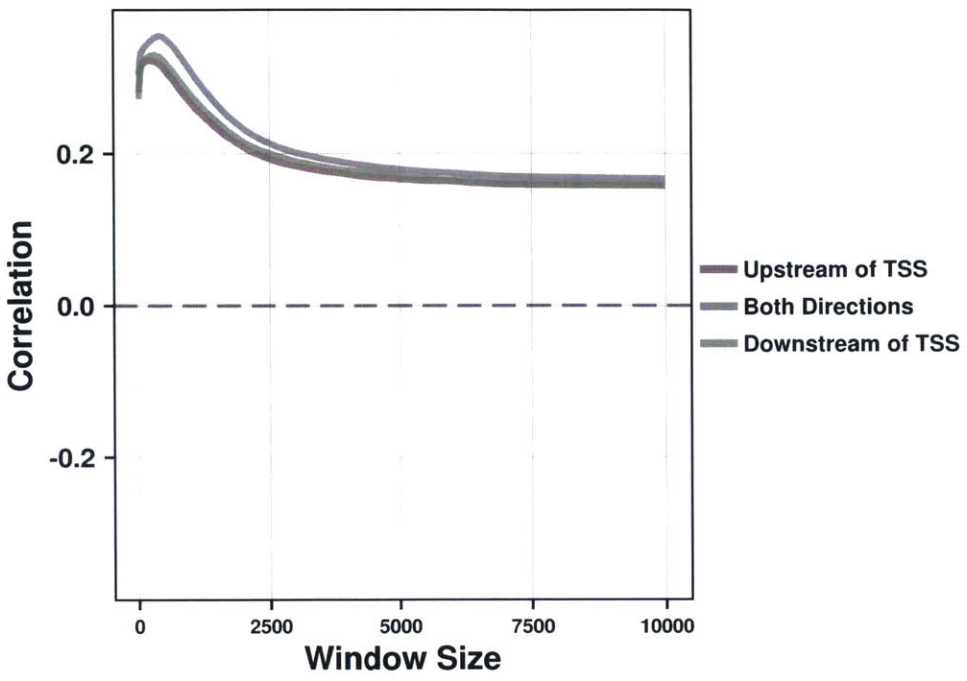

Figure 2-5: Spearman correlation rho between local GC content and transcript expression in the GTEx lung tissue expression subset. Enrichment of GC bases is measured for a range of window sizes extending in both directions from each transcript's start site as well as in the upstream and downstream directions only.

is complicated by the fact that the level of GC-content in the neighbourhood of mammalian genes has been linked to their level of expression (Lercher et al., 2003; Rao et al., 2013). Even when this finding has been disputed, it has been acknowledged that using a TSS window to define proximity to a gene results in a relationship between GC-content in the window and expression (Sémon et al., 2005). This is confirmed in our dataset; there is a weak but consistent correlation between gene expression and local GC content for a range of window sizes extending in both directions as well as only in the upstream and downstream directions from gene TSSs (Figure 2-5).

This implies that whatever expression effects we detect may simply be the result of enrichment of GC bases within the motifs we use lining up with the enrichment of GC bases in the proximity of TSSs. Since we are interested in measuring whether the unique arrangement of bases in each motif can be used to identify binding sites of

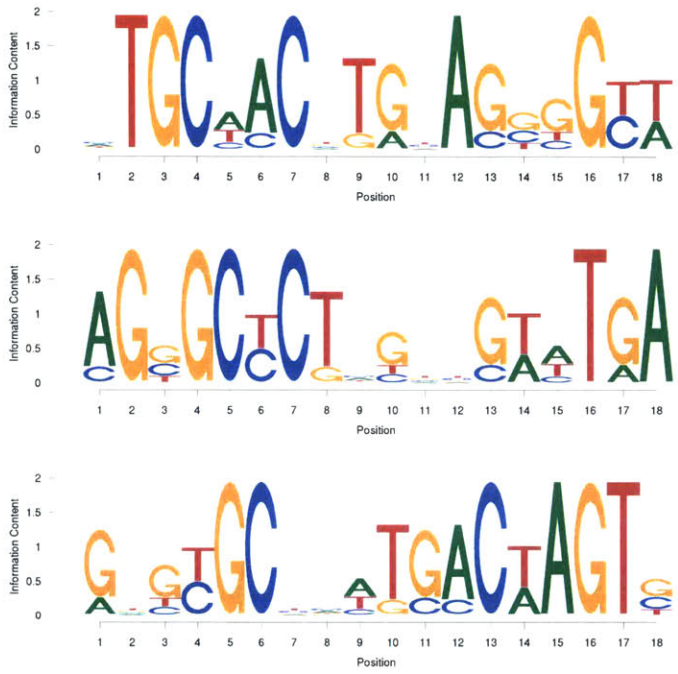

Figure 2-6: Three of the shuffled motifs derived from the original HERPUD1 motif.

some *trans*-factor, we apply a method for correcting for this background expression effect due to GC bias of any particular motif. In particular, we shuffle each of the motifs we use, maintaining the same base composition but randomizing the position of individual bases within the motif, as described in Appendix A.2.

There are nine such "shuffled" motifs available for the HERPUD1 motif, each of which are affected by the same background GC-content biases as the original motif, but none of which should have any ability to predict HERPUD1 binding sites. When we measure the effect on expression of three of these shuffled motifs chosen at random (Figure 2-6), we find that they are similar to those observed for the original motif, and that they are likewise sensitive to changes in the window size we use about TSSs (Figure 2-7). We can thus conclude that this primitive version of the presence model fails to accurately capture regulatory targets of HERPUD1, assuming that *trans*-regulation by HERPUD1 incurs a significant effect on the expression of its targets.

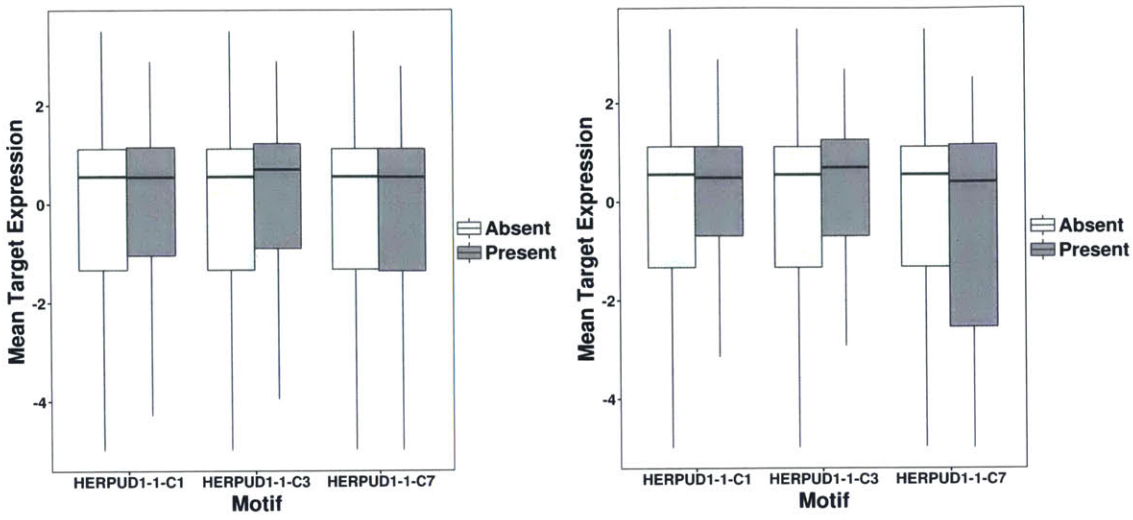

Figure 2-7: First-order expression effects for a ahuffled counterpart of each of the three HERPUD1 motifs shuffled counterparts, using the same motif score and window thresholds to measure regulatory potential.

2.3.1 Improving the Presence Model

Given the failure of this approach, we ask if is possible to tangibly improve upon it using the additional information available in motif hit profiles. The HERPUD1 motif is enriched in the proximity of transcription start sites, both in absolute terms and relative to its shuffled counterparts (Figure 2-8). It hence follows that at least some of this enrichment is due to the tendency of HERPUD1 to bind in the proximity of TSSs, and that this binding has some effect on the expression of nearby transcripts when it occurs, given that HERPUD1 is so abundant in lung tissue. We thus consider more sophisticated ways of measuring the presence of motif hits in the proximity of TSSs to infer regulatory relationships.

For example, can we pick a better threshold than $p = 1 \times 10^{-5}$ for the initial filtering of motif hits? To answer this question, we created a new set of HERPUD1 motif hits than included all hits satistfying much more lax threshold of $p = 0.05$ for both the original motif and each of its shuffled variants. We can then observe the first-order effects of each motif as we vary the threshold used for filtering hits by subsetting

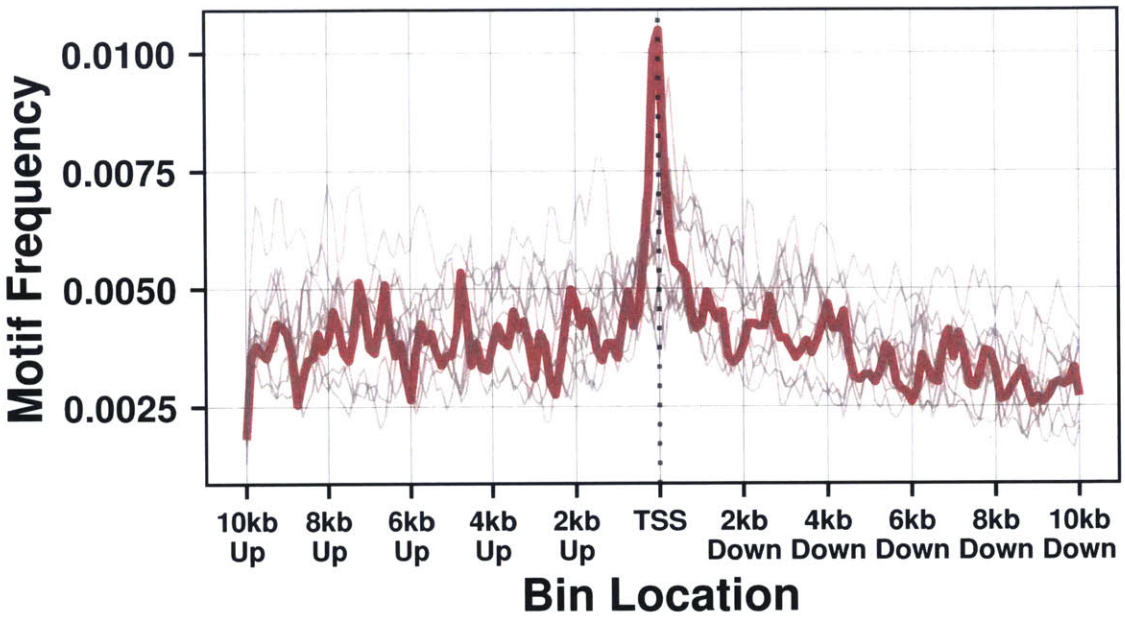

Figure 2-8: The frequency of HERPUD1 motif hits satisfying a $p = 0.05$ score threshold in the proximity of transcription start sites. Frequency is measured using the proportion of transcripts with at least one motif present in each of a series of sliding windows of size 250 base pairs positioned relative to TSSs.

this new set by successively higher motif score thresholds. This allows us to check if the more stringent thresholds for motif scores are indeed appropriate when modelling *trans*-regulation.

At this point we also need a metric to concisely measure first-order expression effects, so that we can compare these effects across a range of motifs and models. We use the Wilcoxon rank sum test of the null hypothesis that a randomly drawn gene linked to a TF will have higher mean expression than a random gene that is not linked to generate a p-value. This p-value is then $\log_{10}$-transformed and given a sign: positive if the genes linked to the TF have higher mean expression, negative otherwise. Thus a signed p-value of 0 indicates no discernable first-order effect, and values above and below 0 indicate an activating and repressing effect on expression respectively.

For example, we can repeat the analysis presented in Figure 2-4, but instead measure the signed p-value for both original and shuffled HERPUD1 motifs across a range of motif score thresholds (Figure 2-9). Clearly, simply incorporating lower-scoring motif hits into the presence model does not improve its performance. For window sizes of both 5kb and 1kb about TSSs, no threshold results in a statistically significant difference in the expression of transcripts with proximal qualifying motif hits. Although some shuffled motifs are able to produce such a difference at the larger window size and lax motif thresholds, this is most likely due to the GC-bias discussed previously.

However, we must also question whether using window sizes measuring in the thousands of base pairs about transcription start sites is an appropriate way of assessing the potential targets of HERPUD1. As we have already seen in Figure 2-8, the enrichment of the HERPUD1 motif is confined to a narrow region around TSSs; upon closer examination, this region is no greater than 300 base pairs in width, and is centred slightly upstream of TSSs (Figure 2-10). Although motif hits located outside of this window may still be associated with regulatory activity, it seems likely that

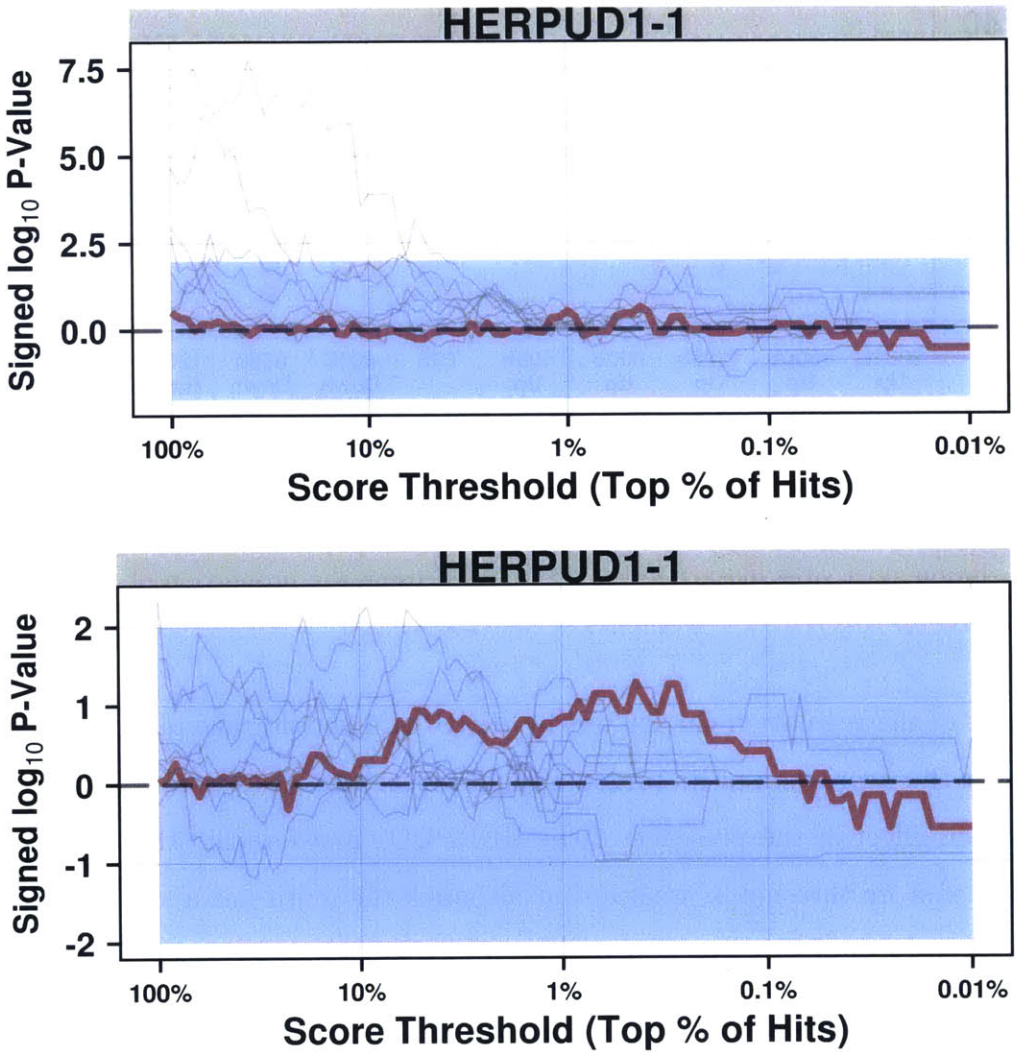

Figure 2-9: First-order expression effects, as measured by Wilcoxon rank-sum signed p-value, for the HERPUD1 motif and its corresponding shuffled motifs across a range of possible motif score thresholds. Transcripts are partitioned using motif presence within 5kb of TSSs (top) and within 1kb (bottom).

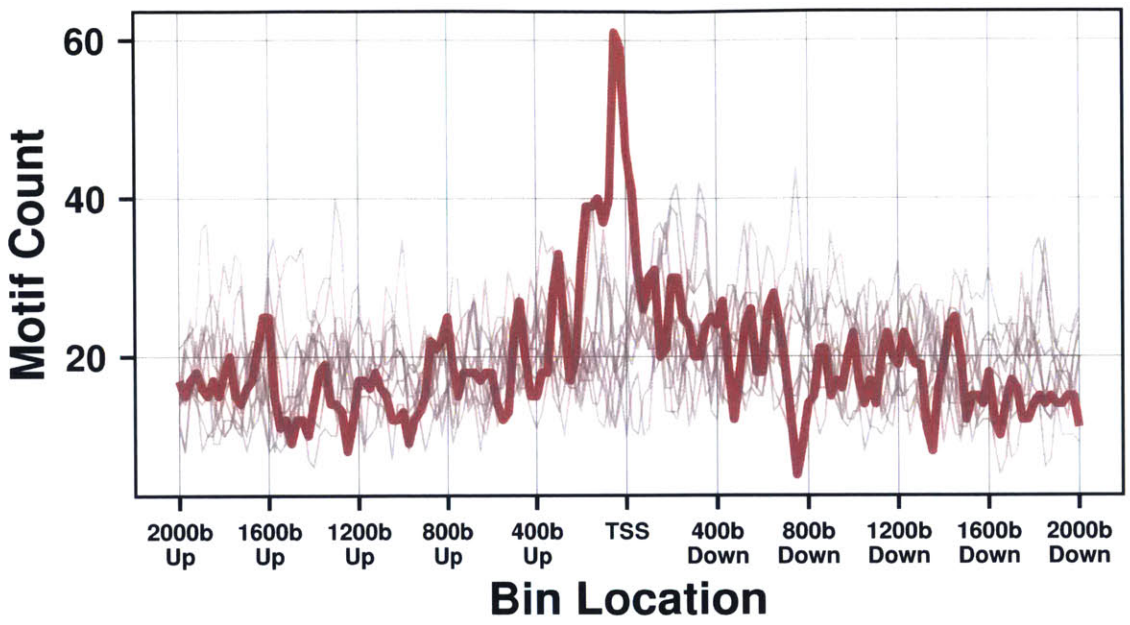

Figure 2-10: The frequency of HERPUD1 motif hits in the immediate proximity of transcription start sites, with frequency measured using the proportion of transcripts with at least one motif present in windows of size 50bp.

the bulk of the regulation carried out by HERPUD1 should take place in this region, and that the expression of its targets should reflect this. However, When we test the expression effect of the presence of the HERPUD1 motif within this window across a range of score thresholds, we see that although the motif has a stronger repressive effect than its shuffled variants at low score thresholds, this effect is not statistically significant (Figure 2-11).

Nevertheless, when we measure the expression effects described above using $\log_{10}$ fold-change in expression, we see that although the original HERPUD1 motif does not partition transcripts according to expression level any better than its shuffled variants at the larger window sizes, it does do so at the narrow window when all motif hits are included (Figure 2-12). Furthermore, the advantage of using signed p-values in this context is demonstrated, as the fold-change metric becomes unstable at strict motif score thresholds when relatively few transcripts are identified as regulatory targets.

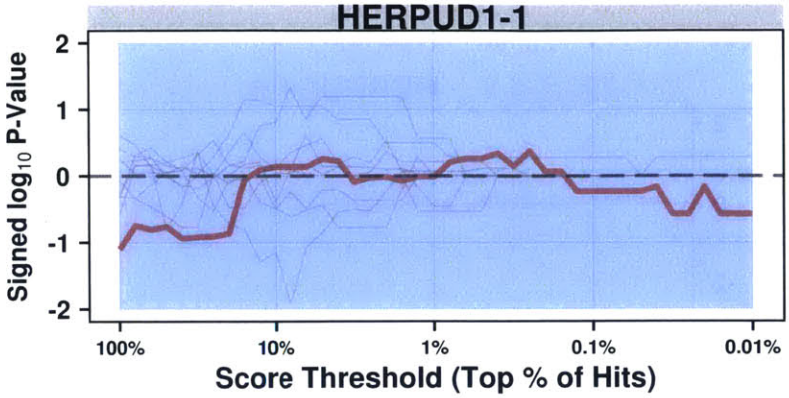

Figure 2-11: Signed p-value first-order expression effects for the HERPUD1 motif and its corresponding shuffled motifs across a range of possible motif score thresholds and a TSS window filter of $[-200bp, 100bp]$.

To support the use of smaller window sizes to gage HERPUD1 regulatory activity, we can vary the size of the TSS window filter instead of the motif score threshold to see what effect this has on the performance of the presence model (Figure 2-13). We see that the original HERPUD1 motif produces a significantly better partition of transcripts than its shuffled variants only at very small window sizes of less than 200 base pairs. This suggests an interaction between the presence of HERPUD1 motif hits and promoter regions in the proximity of TSSs as is suggested by the enrichment of motif hits very close to TSSs.

Having described how to check the direct effects of *trans*-regulation through the presence of motif hits in the vincinity of a gene's TSS on its expression, we now expand on these models to include the coexpression of the transcription factor in question and target genes. Whereas we previously calculated the mean expression of each gene in our dataset and then compared the expression of genes linked to a TF by a given motif hit model to the rest, we now repeat the same analysis substituting Spearman coexpression with the transcription factor in question for mean expression. We exclude the bottom 50% of genes by mean expression in order to avoid having lowly-expressed genes with noisy expression profiles confound our results.

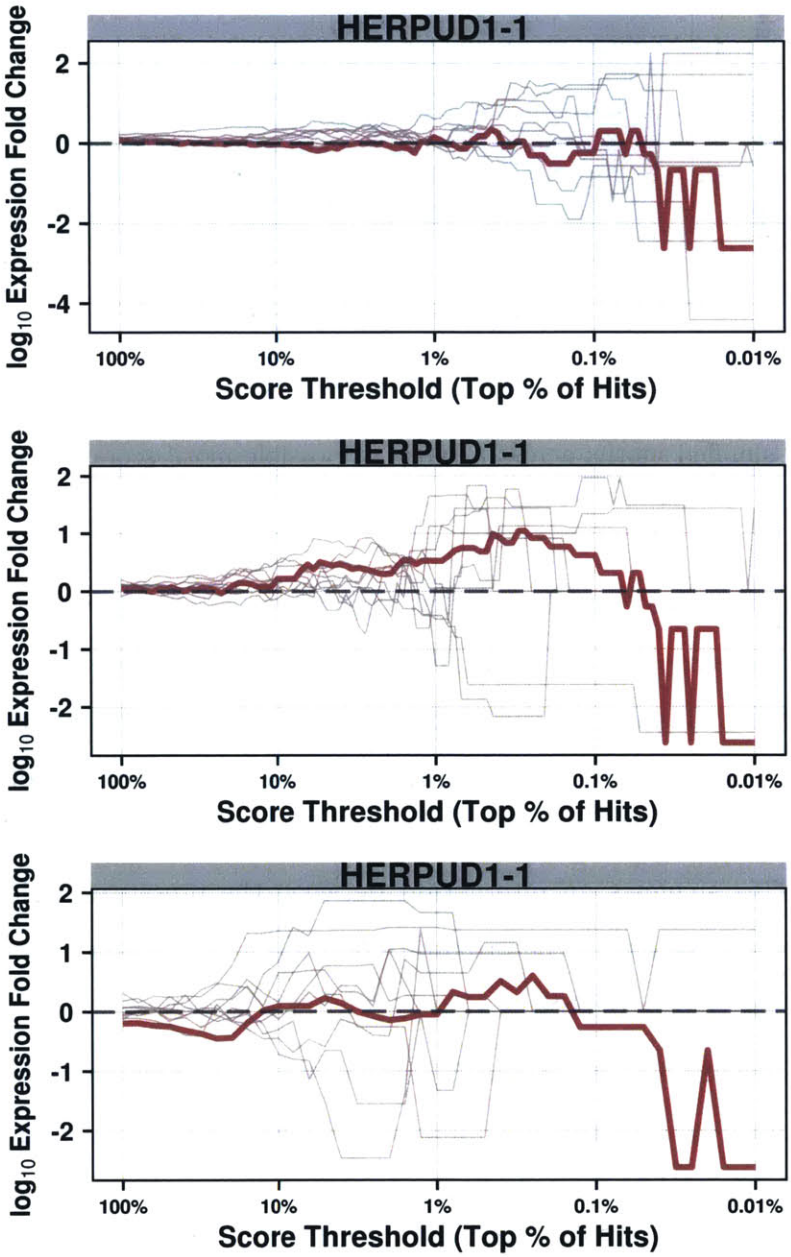

Figure 2-12: First-order expression effects, as measured by expression fold change, for the HERPUD1 motif and its corresponding shuffled motifs across a range of possible motif score thresholds. Transcripts are partitioned using motif presence within 5kb of TSSs (top), within 1kb of TSSs (middle), and within a $[-200bp, 100bp]$ window around TSSs (bottom).

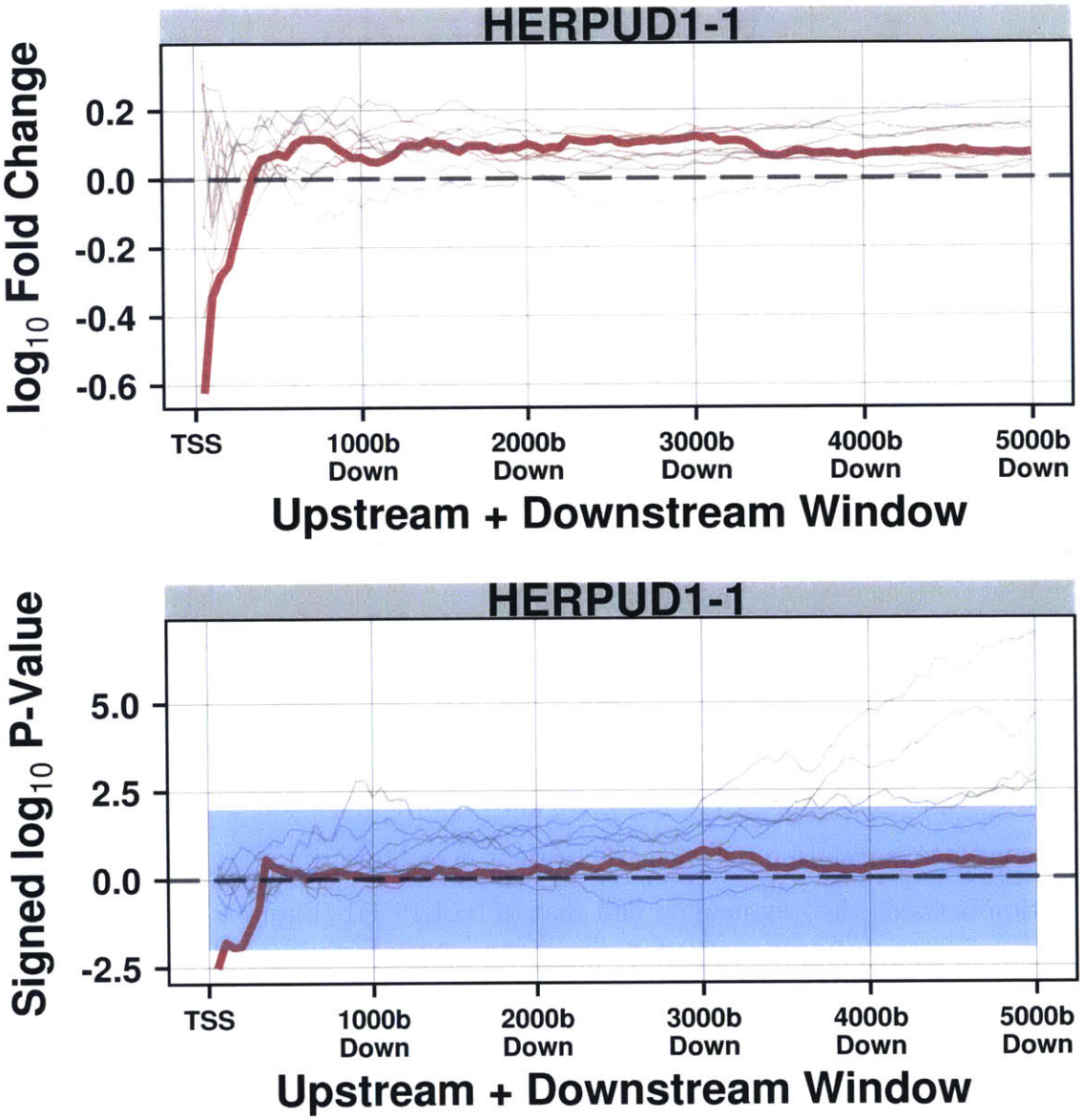

Figure 2-13: First-order expression effects across a range of possible window sizes as measured using log fold-change (top) signed p-values (bottom).

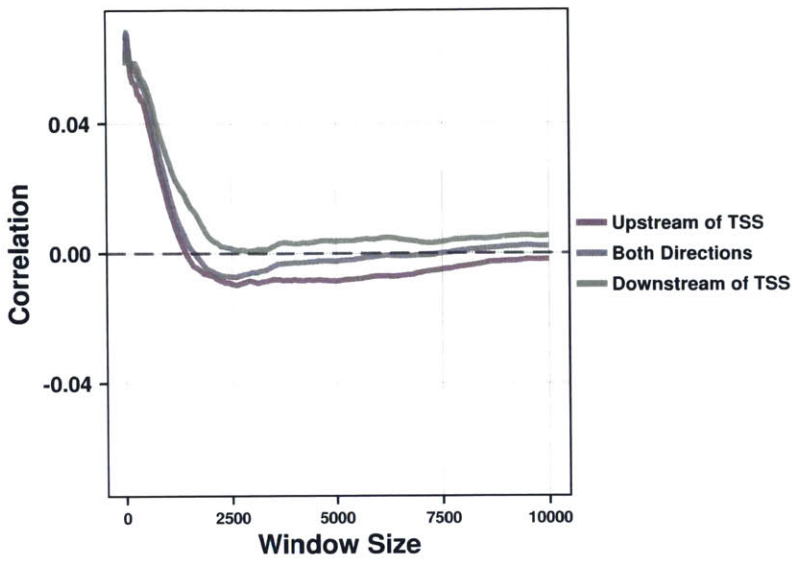

Figure 2-14: Spearman correlation rho between local GC content and transcript coexpression with HERPUD1 in the GTEx lung tissue expression subset.

We can also use similar metrics to evaluate the performance of our hit models, with a signed p-value being calculated from Wilcoxon rank sum test comparing coexpression values between the two sets of genes and a simple difference between the mean coexpressions taking the place of expression fold change. Because we see that there is some correlation between GC content near transcript start sites and the correlation between their expression and that of HERPUD1 (Figure 2-14), we again use the hits of shuffled motifs to correct for background nucleotide bias.

As with first-order expression effects, we see that varying motif score thresholds does not result in a partitioning of target transcripts that differentiates between transcripts coexpressed with HERPUD1 and those that aren't (Figure 2-15). Using all available motif hits and varying the size of the window about the TSS does not result in well-performing partitions either (Figure 2-16). We are thus left to conclude that using motif hits in isolation in the presence model is insufficient to identify regulatory targets of HERPUD1 as measured using effects on coexpression.

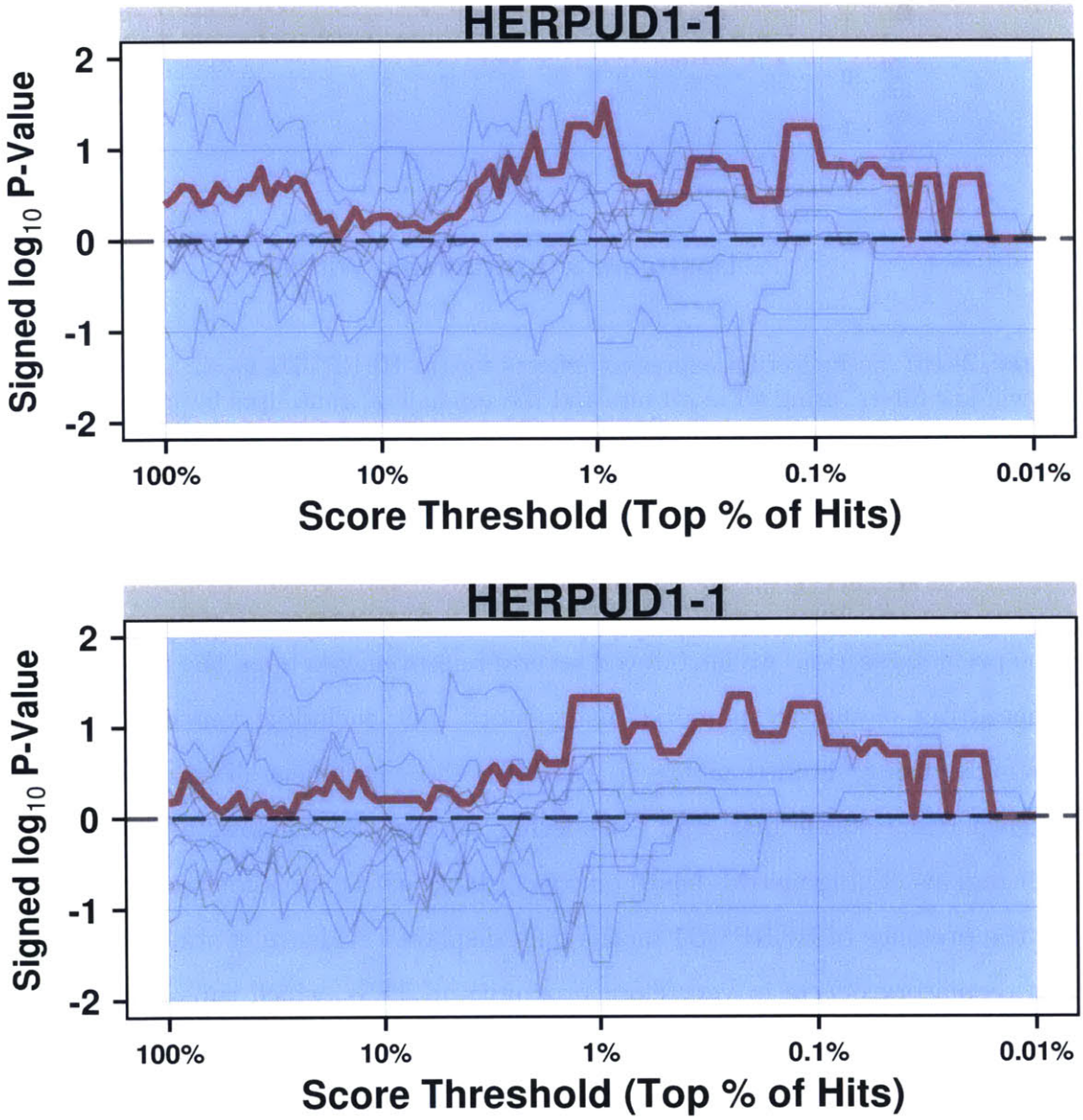

Figure 2-15: Second-order expression effects, as measured by signed p-values, for the HERPUD1 motif and its corresponding shuffled motifs across a range of possible motif score thresholds. Transcripts are partitioned using motif presence within 5kb of TSSs (top) and within 1kb (bottom). Only the top half of transcripts by mean expression are used to measure coexpression effects.

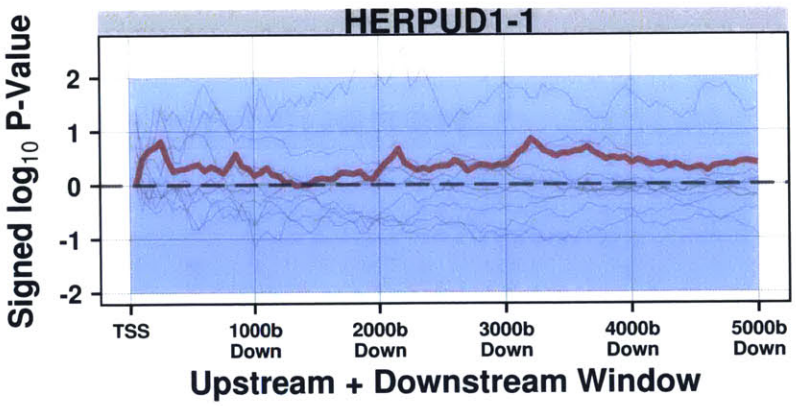

Figure 2-16: Second-order expression effects for the HERPUD1 motif across a range of TSS window filters, using all motif hits and the top half of transcripts by mean expression.

2.4 Constructing Motif Hit Models for Various Transcription Factors

In the previous section, we introduced several improvements upon the presence model of measuring regulatory potential using motif hits, including varying motif score thresholds, using shuffled motifs to establish the enrichment of motifs relative to background nucleotide bias, and considering various TSS windows as proximal promoter regions. Using metrics based on concordance with expression profiles, we found that the presence of HERPUD1 motifs only displayed evidence of statistically significant regulatory effects in very close proximity to TSSs, which runs counter to the wide windows often used by previously described models of *trans*-regulation. We now extend this analysis to other transcription factors active in lung tissue, and consider various improvements on the presence motif hit model.

For instance, ATF4 (activating transcription factor 4) is the most highly expressed TF for which motifs are available (Figure 2-17). Although the three motifs available for ATF4 are fairly similar to one another, they each have a unique profile of enrichment in the proximity of TSSs (Figure 2-18). In particular, the ATF4-1 motif occurs more frequently around TSSs, but not much more frequently than its shuffled

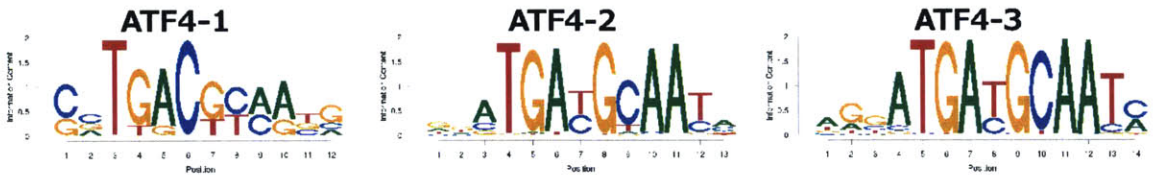

Figure 2-17: The three motifs associated with ATF4.

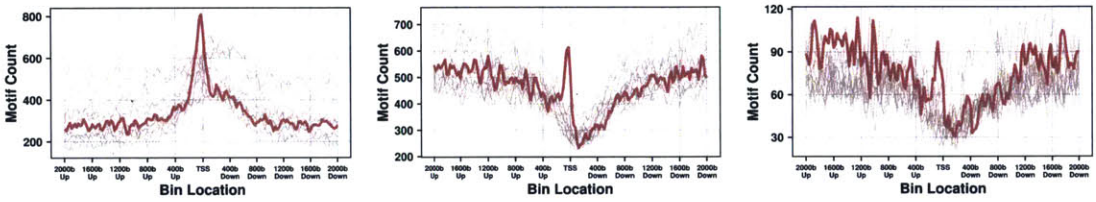

Figure 2-18: The frequency of ATF4 motif hits in the proximity of transcription start sites, with frequency measured using the proportion of transcripts with at least one motif present in windows of size 50 base pairs.

counterparts, ATF4-2 becomes depleted closer to TSSs but is enriched even compared to its shuffled variants right at TSSs, and ATF4-3 becomes slightly depleted close to TSS but not more so than its shuffled versions.

Up until this point we have followed the lead of previous work in this area, and considered all motif hits satisfying some score threshold to be equal to one another. However, this approach neglects an important source of information, namely the scores of the motif hits that remain after the threshold has been applied. Using the expression-based methods we have developed for validating regulatory relationships effects we can examine the benefits of using models that, unlike the presence model, incorporate motif hit scores when predicting targets of TFs.

We will start with a "sum" model, in which all available motif hits in a certain window around the TSS are used, and their hit scores are summed to create a regulation score for each gene. By choosing some cutoff for the regulation score, analogous to using a threshold for motif hit scores for the presence model, we can calculate expression effects of motif hit profile models as before.

There are several reasons why this sum model should be superior to the presence

model. By taking in as much available data on motif hits before applying a cutoff, we get a much richer picture of the genomic sequence patterns in the vincinity of each TSS that might contribute to TF binding. Whereas the presence model throws away all motif hits below a threshold before calculating a presence score (in effect, a binary variable), the sum model uses all motif hits to calculate a continuous presence score that can then be thresholded as necessary. This allows for more subtle patterns of TF binding to be incorporated into our regulatory model.

For instance, some TFs may be more inclined to bind to areas with a large number of low-scoring motif hits as opposed to a small number of high-scoring hits. Under the presence model, the former may be considered as devoid of regulatory activity, depending on the motif score threshold used, while the sum model would allow for a sufficiently large number of weak motif hits to score better a few strong hits in a given promoter region. Furthermore, a TSS that is surrounded by exactly one high-scoring hit is given exactly the same score as a TSS that is surrounded by many high-scoring hits in the presence model, when we would expect the latter to be much more likely to attract TF binding, which is concordant with the sum model.

On the other hand, by using information from low-scoring hits, the sum model may be confounded by a relatively noisy source of data. Low-scoring hits may be much more likely to be caused by background biases in nucleotide composition than some sequence that is actually relevant for TF binding. The composition of some TF proteins may be such that they ignore sequences corresponding to low-scoring hits entirely when binding to the DNA molecule. Nevertheless, because the sum model weighs motif hits according to their score, these effects, if present, should at very least be ameliorated. We can also apply the sum model to the shuffled variants of each motif to correct for this background nucleotide bias.

The use of the sum model is further supported by the enrichment of ATF4 motifs in the proximity of TSSs when the sum of motif hit scores in a given area is used to measure the density of motif matches as opposed to frequency of matches. As

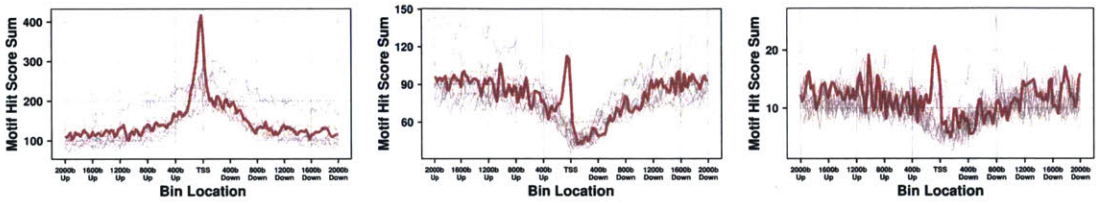

Figure 2-19: The sum of ATF4 motif hit scores within windows of size 50bp in the proximity of TSSs.

shown in Figure 2-19, using hit score sums produces higher peaks of enrichment in the proximity of TSSs for the three original ATF4 motifs relative to their shuffled counterparts, especially for ATF4-1 and ATF4-3. This indicates that not only do ATF4 motifs occur more frequently close to TSSs, the matches that occur proximal to TSSs are of higher confidence, suggesting greater potential for binding of the ATF4 protein.

As was the case with HERPUD1, the enrichment profiles of the ATF4 motifs measured using motif presence and motif score sums suggest that regulatory binding occurs in a narrow window around TSSs. However, this window seems to vary slightly between the three motifs; roughly $[-200b, 100b]$ for ATF4-1, $[-150b, 100b]$ for ATF4-2, and $[150b, 0]$ for ATF4-3. The presence model shows some ability to distinguish regulatory targets of ATF4 in the case of ATF4-1 compared to background nucleotide bias for any given motif score threshold, but this performance quite weak and is highly sensitive to the motif score threshold used (Figure 2-20). Using the sum model results in a a more robust effect on expression as measured using the ATF4-1 motif, though it does not result in an improvement in expression effects for ATF4-2 and ATF4-3 (Figure 2-21).

To facilitate the direct comparison of different motif profile models, we introduce now normalize expression scores for motifs relative to their shuffled counterparts. In the case of signed p-values, this is done by subtracting the mean signed p-value obtained using each of shuffled variants from that obtained using the original motif

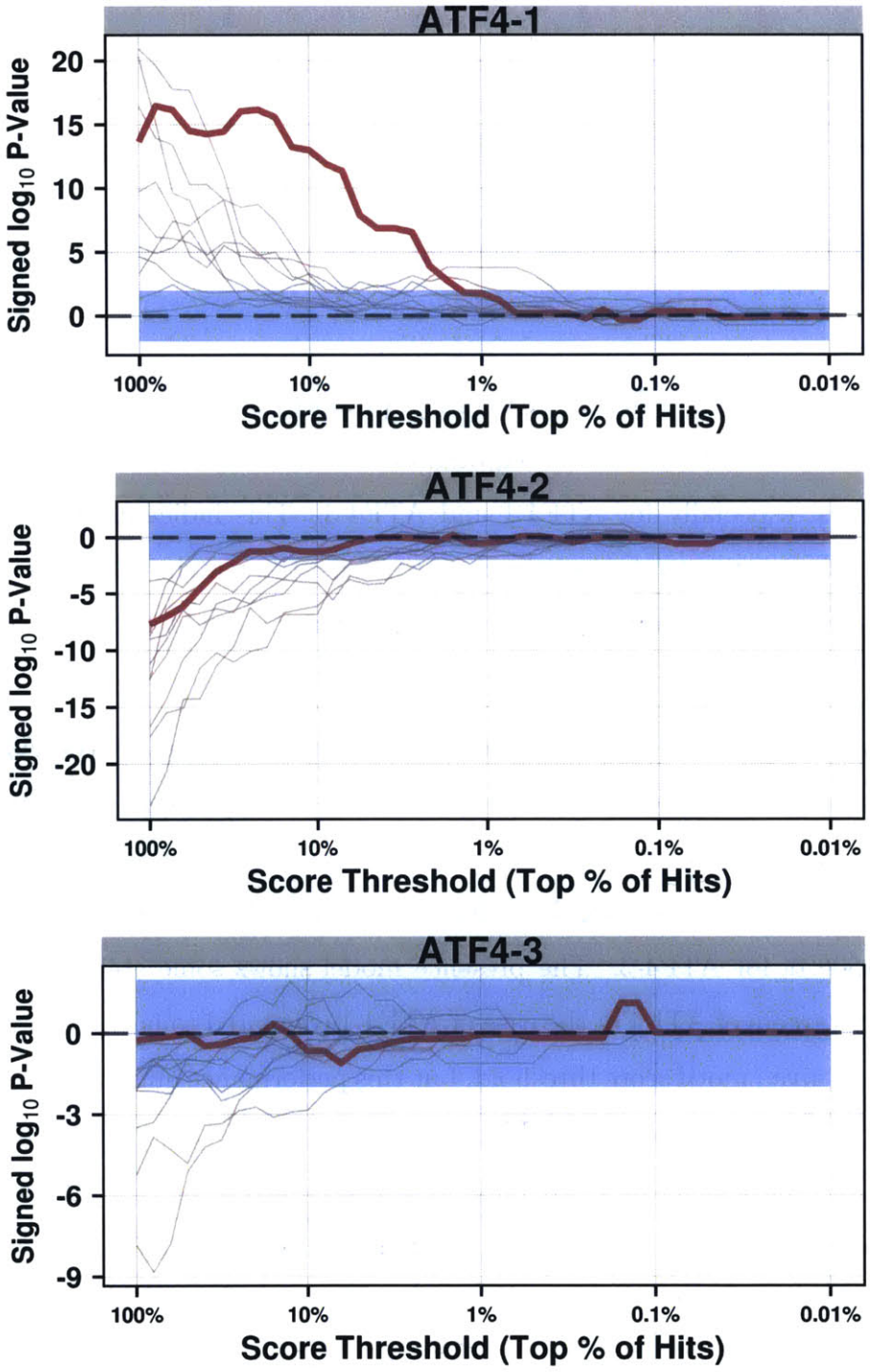

Figure 2-20: First-order expression effects of ATF4 using various motif score thresholds for ATF4-1 with a TSS window of $[-200b, 100b]$ (top), ATF4-2 with a window of $[-150b, 100b]$ (middle), and ATF4-3 with a window of $[150b, 0]$ (bottom).

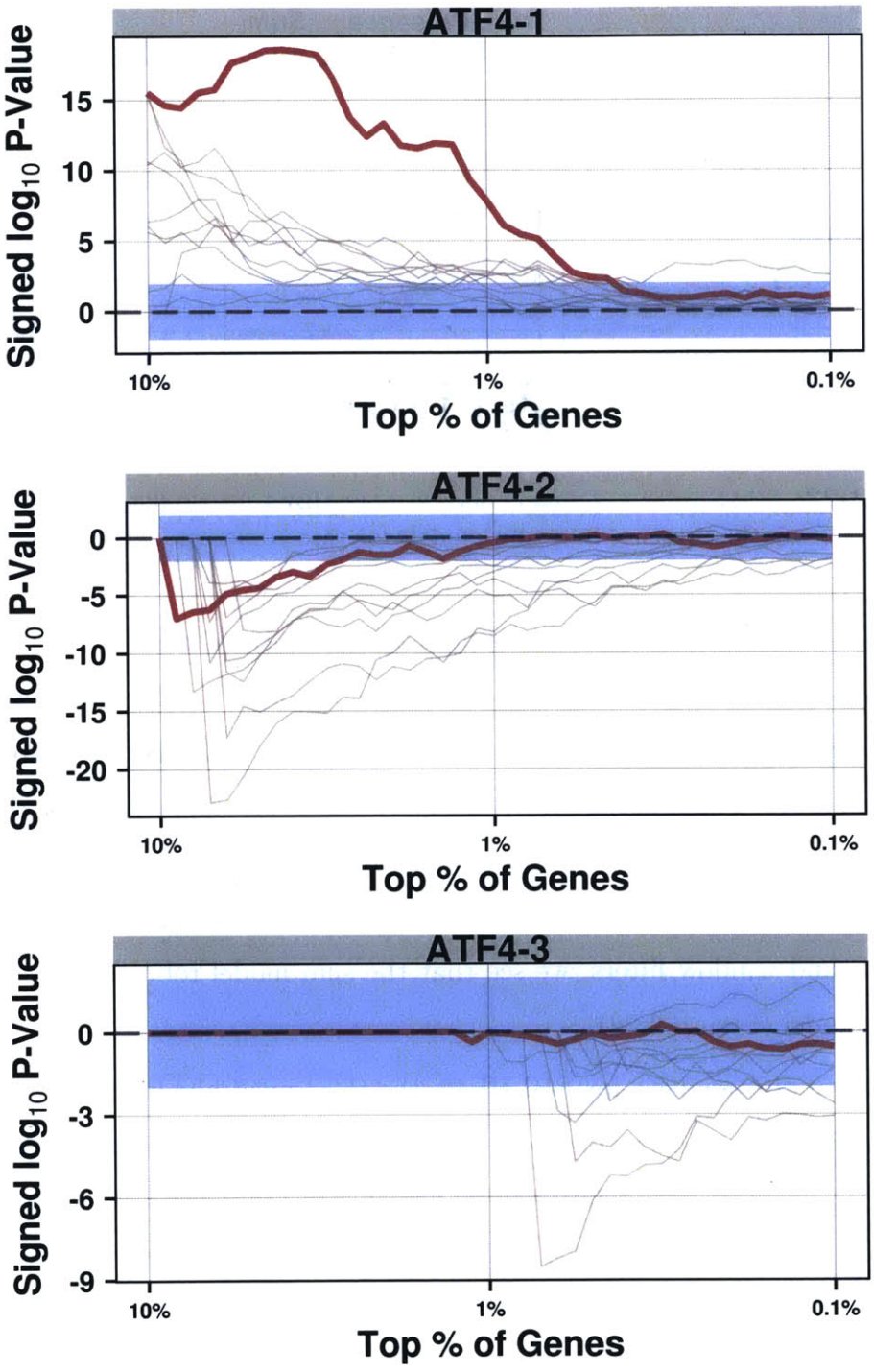

Figure 2-21: First-order expression effects of ATF4 measured using the sum motif hit profile model using various model score cutoffs for ATF4-1 with a TSS window of $[-200b, 100b]$ (top), ATF4-2 with a window of $[-150b, 100b]$ (middle), and ATF4-3 with a window of $[150b, 0]$ (bottom).

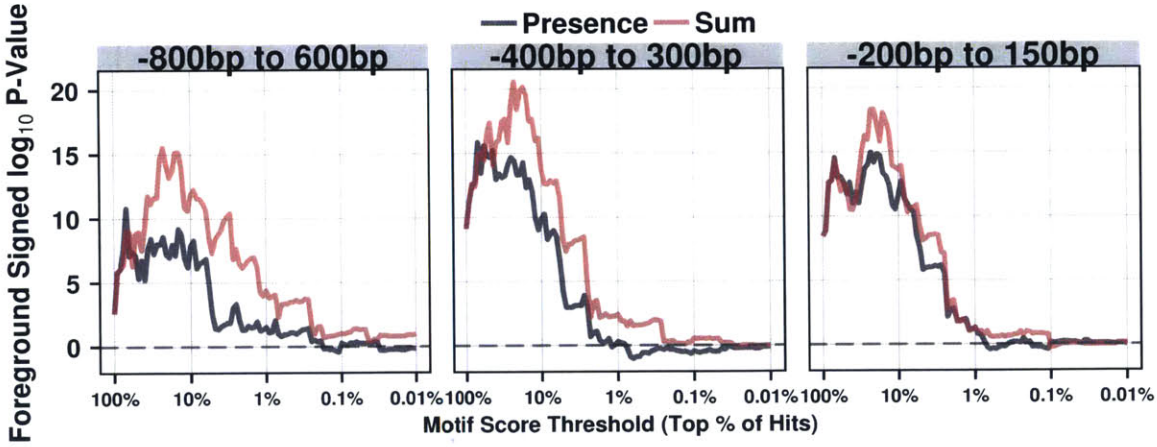

Figure 2-22: First-order expression effects, normalized for shuffled motif model scores, for ATF4-1 using the presence and sum models for three different TSS window filters.

for each model tested. Because the average presence of the shuffled variants for each motif constitutes a background score for how likely it is for the motif to match the given sequence based on nucleotide bias alone, we call this normalized metric a foreground score. For example, if we calculate the foreground scores for the presence and sum models for the ATF4-1 motif for a range of thresholds (normalized between the two models such that the same number of genes are identified as targets at each threshold) and window filters, we see that the sum model robustly outperforms the presence model (Figure 2-22).

We repeat this analysis for BHLHE40, the third-most highly expressed transcription factor in the lung tissue dataset. We see that the motif for BHLHE40 is enriched at transcription start sites relative to its presence in other areas of its genome, but is in fact depleted in absolute levels relative to its shuffled variant, and that this effect is once again stronger when the sum of motif hit scores is taken into account rather than just motif presence (Figure 2-23). Also, this peak is heavily skewed towards the upstream side of TSSs, and is somewhat broader than the ones we observed for the ATF4 and HERPUD1 motifs, measuring roughly 400 base pairs in width. This underlines the need to apply motif hit profile windows that are unique to each transcription

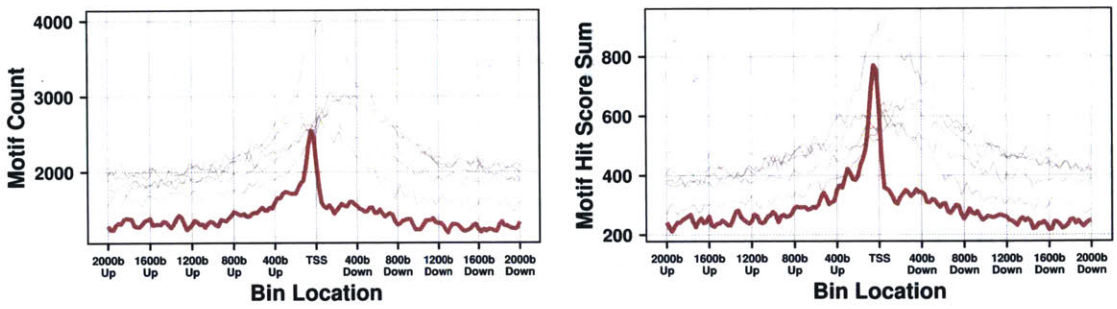

Figure 2-23: The frequency (left) and motif hit score sums (right) of BHLHE40 motif hit scores within windows of size 50bp in the proximity of TSSs.

factor, rather than using the same profile model for all available TFs.

When we apply the presence and sum models to the profiles of the BHLHE40 motif using window filters that correspond to these enrichments, we see that the sum model does not produce an improvement in performance over the presence model (Figure 2-24). This suggests that the weakly-scoring motif hits for BHLHE40 include enough background noise that they overwhelm the power of stronger hits to predict binding which influences expression of target genes. It is hence clear that using a single motif hit profile function such as score sums may not be appropriate for all transcription factors.

There is another wrinkle that we can add to the sum model: weighing motif hits according to their distance to the TSS. Based on the general properties of motif hit enrichment profiles in the proximity of TSSs, it seems likely that hits that are located further away from TSSs should have a smaller likelihood of influencing expression, all other things being equal. We thus add a multiplicative weighing factor when computing sums of motif hit scores based on the window size that deprecates linearly. For example, a hit located 200 base pairs from a TSS under a model using a window of size 400bp has its motif hit score multiplied by 0.5, a hit located 100 base pairs from a TSS under a model using a window of size 500bp has its motif hit score mutiplied by 0.8, and so on.

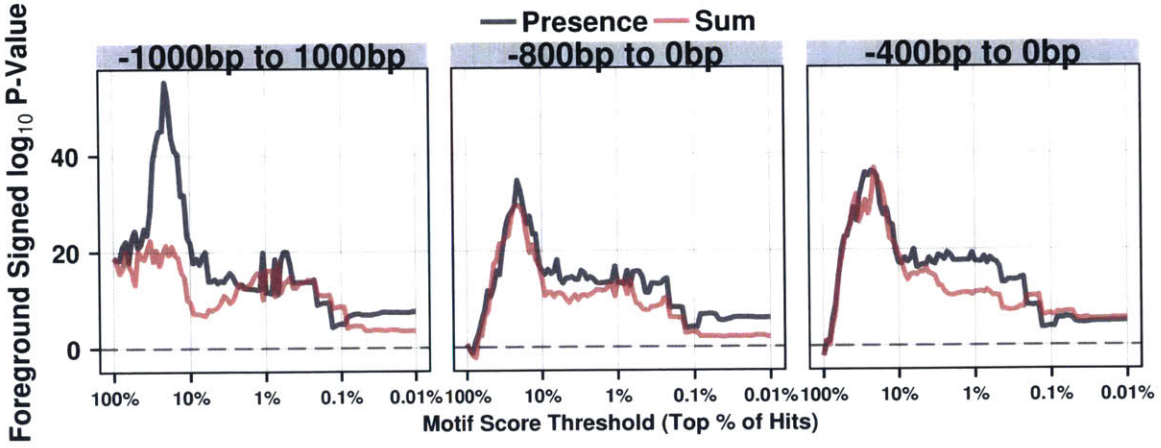

Figure 2-24: Normalized first-order expression effects for BHLHE40 using the presence and sum models for three different TSS window filters.

As shown in Figure 2-25, when we compare the performance of this linear distance model, the sum model, and a model whose score is simply the count of motif hits occuring within a window for BHLHE40 and the a motif for ELF3, we once again observe the heterogeneity inherent between the expression effects of different motif profile scoring functions. The count model performs very well for BHLHE40 relative to the other three models, which implies that motif hit scores don't have any relation to binding occupancy for its motif. On the other hand, the linear distance model outperforms the other three models in the case of the ELF3-1 motif, especially for larger window sizes, demonstrating its ability to correctly weigh less proximal hits to gage regulatory potential.

Before we test these models on any more transcription factors, we consider what other improvements we can make on the presence and sum models, and how we can fit them into a general framework for constructing motif hit profile functions.

2.4.1 Introducing Motif Profile Functions

We have thus far presented a framework for modeling *trans*-regulation using the presence of transcription factor motif hits in the proximity of gene transcription start sites,

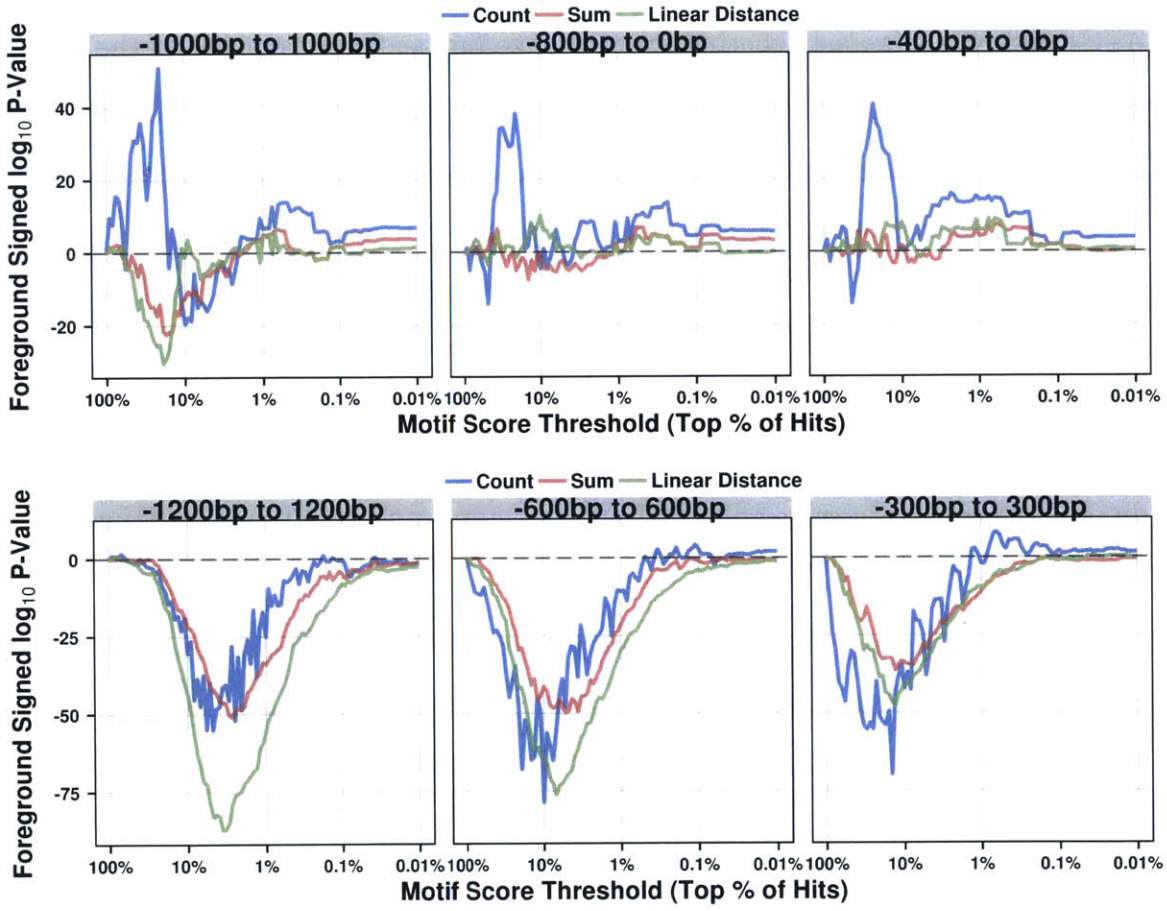

Figure 2-25: Normalized first-order expression effects relative to the effects calculated using the presence model for a range of TSS window filters for the motif count, motuf sum, and linear distance models for the BHLHE40-known3 motif (top) and the ELF3-1 motif (bottoms).

as well as a series of metrics for gauging the performance of such models when compared to an expression dataset. We have also introduced a simple way of expanding upon this model, by using a gene score cutoff based on the sum of all hits near each TSS in place of a motif score cutoff. Given that this sum approach yielded improvements in performance when modeling the *trans*-regulation of several TFs, including ATF4 and ELF3, we ask whether we can make further improvements to the model.

An important difference between the *regulatory*-models used by Neph et al. and Glass et al. is the size and shape of the TSS window filter they applied to motif hits. There is a general agreement that motif hits closer to the TSS indicate a stronger regulatory relationship, as supported by the findings of Pique-Regi et al., but there exists discord as to just how close hits have to be to the TSS for their effect to be significant, and whether hits upstream of the TSS are more significant than those downstream of the TSS. Moreover, can we expect these regulatory properties of motif hits to hold constant across all transcription factors and tissue types, and if not, what kinds of variation exist?

The first observation we make is that the use of window filters is a very crude way of putting greater regulatory emphasis on motif hits occuring closer to TSSs, as it is highly unlikely that a motif hit fifty base pairs beyond some threshold distance will have no effect on regulation relative to an identical hit located fifty base pairs within the threshold. Pique-Regi et al. recognize this by using the inverse of the distance to the nearest TSS as a multiplicative factor when calculating the prior probability that a given site is bound by a TF. Are other distance functions a better fit, however, for modeling how much of an effect a given motif hit will have on the regulatory relationship between its associated TF and the nearest gene? Using our expression effect metrics we can now thoroughly test this hypothesis.

Furthermore, is it possible that the presence of many weakly significant motif hits in the proximity of a TSS is a better indicator of regulation than the presence of a single strongly significant hit? How important is the score a motif hit in de-

termining its regulatory effect? Is it possible that the directionality of a hit (i.e. on the sense or anti-sense strand relative to the gene) should also be taken into account? How strongly should we weigh the overlapping presence of other sources of genetic and epigenetic information, such as evolutionary conservation, DNase1 cleaving, histone marks, and the presence of other *trans*-regulatory elements that may bind co-operatively or competitively to the TF in question?

To answer these questions, we construct a model of *trans*-regulation that can take all of these factors into account, with parameters corresponding to our belief in how the hits of a given transcription factor tend to be situated in the proximity of genes that the TF regulates. At the core of our model are three families of functions that describe how a profile of putative binding sites associated with a TF in the proximity of a particular gene's transcription start site affect its transcription. The goal is to produce a continuous value for each (TF, gene) pair that describes how strongly the TF affects the expression of the gene.

We first consider the role that the distance between a motif hit and the target gene's transcription start site plays in our model. A distant binding site should have a smaller effect on transcription, all other things being equal; we thus examine decreasing functions over distance. We allow our model to choose whether this function should decrease at an increasing rate or decreasing rate using a convexity parameter $c>0$, hence allowing for various levels of sensitivity of regulation to distance. We also choose a window size $w>0$ that defines the maximum distance over which a binding site has any effect on transcription. Although this may reminiscent of the window filter that was disparaged above, because our distance function allows regulatory effect to taper off gradually over distance, we avoid the pitfalls associated with choosing an arbitrary distance threshold at which regulatory effect is abruptly terminated.

The core distance function used to calculate the transcriptional effect of a single

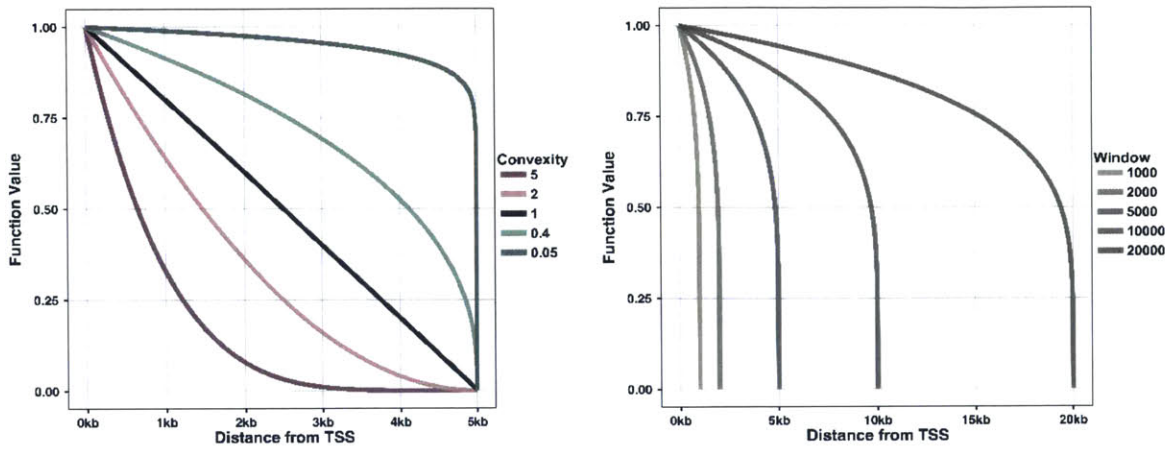

Figure 2-26: Model functions for various values of the convexity (left, with window at 5000) and window size (right, with convexity at 0.2) parameters.

binding site based on its distance to the target gene's TSS d_i is thus:

$$\mathcal{D}(d_i;\, w, c) = \begin{cases} \left(\frac{w-d_i}{w}\right)^c & \text{if } w > d_i \\ 0 & \text{otherwise} \end{cases}$$

As shown in Figure 2-26, the window and convexity parameters endow this model with flexibility for a range of different interactions between binding distance and regulation. Some TFs may be agnostic to the distance from the TSS as long as they are close to it, which would be modeled by a $\mathcal{D}$ with a convexity close to zero and small window size. Other TFs may be able to influence transcription through binding sites far away from a TSS but at a much smaller rate than closer binding sites. This behaviour is captured by a $\mathcal{D}$ with a convexity equal to one for a linear depreciation of influence over distance, or a convexity greater than one for polynomial depreciation.

After distance to the nearest TSS, the other fundamental property of an individual motif hit is its score $0 < s_i \leq 1$ as determined by the PWM that generated it. As discussed above, we used a very low threshold for calling motif hits, and so we have access to a greater range of weakly matching motif hits as well as strongly matching motif hits. Nevertheless, it may be possible that using motif hits with low scores is a

detriment, and so we design a function that can take this into account:

$$\mathcal{S}(s_i;\, \varsigma) = s_i^{\varsigma}$$

The parameter $\varsigma \geq 0$ thus corresponds to how strongly we weigh high-scoring motif hits to low-scoring motif hits: at $\varsigma = 0$ all hits are weighed the same according to score, and for increasing values of ς scores closer to the maximum of one are given greater weight. This can be shown mathematically by considering the derivative of $\mathcal{S}$ with respect to ς: $s_i^{\varsigma} \log(s_i)$ has uniformly larger values for s_i closer to zero.

Having chosen distance and sum functions for individual motif hits, we now consider how to aggregate the profile of motif hits occuring in the proximity of each TSS into a single value. We have already implicitly made this decision in the presence and sum models discussed previously: the former used the maximum function for summation while the former used the arithmetic sum, with both using the identity function $f(s_i) = s_i$ for hit scores and the signal function $f(d_i; w) = 1$ if $d_i \leq w$, 0 otherwise for hit distances.

For a given gene, we have the profile of n hits in the vincinity of its TSS, each defined by its distance to the TSS and motif Ascore: $H = \{(d_1, s_i), (d_2, s_2), \ldots, (d_n, s_n)\}$. For each hit, we multiply its distance and score function to get a score h_i. To aggregate these scores into a final score for the (TF,gene) pair, we propose the function

$$\mathcal{E}(H;\, \sigma) = \left(n^{\sigma} \sum_{i=1}^{n} h_i\right)$$

for $\sigma \geq 0$ taking some real value. We see that $\sigma = -1$ corresponds to the arithmetic mean of all hit scores, and that $\sigma = 0$ corresponds to taking their sum. The advantage of using a continuous parameter σ lies in our ability to fine-tune how much we want to emphasize the importance of quality hit scores in rating a profile of motif hits as opposed to the quantity of hits. For high values of σ, the number of motif hits dominates the sum function, and so we don't care about how strong motif hits are or

where they are placed in relation to the TSS - we only care that there are a lot of them. For values of σ close to zero, the number of motif hits loses importance, as a high final value will only be achieved if the average score assigned to each hit is high. This in conjunction with the flexibility inherent in our distance and scoring functions allows for a wide range of different regulatory binding profiles to exist between transcription factors and their regulatory targets.

2.4.2 Testing Advanced Motif Profile Functions

Having introduced a more general version of the sum and linear distance motif hit profile functions whose performance we have already investigated, we now see if we can better identify regulatory targets of transcription factors using this framework. For example, in the case of the ATF4-1 motif, which discriminated genes according to expression levels significantly better under the sum model for larger window sizes, varying convexity values do not add much in the way of potential in terms of predicting highly-expressed genes. However, for other motifs, such as CEBPB-disc1, we see that high convexity values lead to consistently better expression effects (Figure 2-27).

We can further test various settings of the sum and score parameters on different transcription factors to see whether sensitivity of expression effects to the weighing of motif hit scores and motif hit counts varies between factors. As shown in Figure 2-28, this is indeed the case, as the profile function as applied to the ATF4-1 motif exhbits peak performance when the at a sum parameter of 1 and high score parameters, especially for large window sizes. This suggests that the best way to infer the regulatory targets of ATF4 is to weigh in favour of genes will many ATF4-1 motif hits proximal to their transcription start site, with particularly large weight given to genes with a large number of high-scoring hits.

However, this is not the case for all transcription factors, as the performance of these functions for the BHLHE40-known3 and CTCF-known1 motifs is considerably different. In the case of the former, low score and sum parameters fare better across a

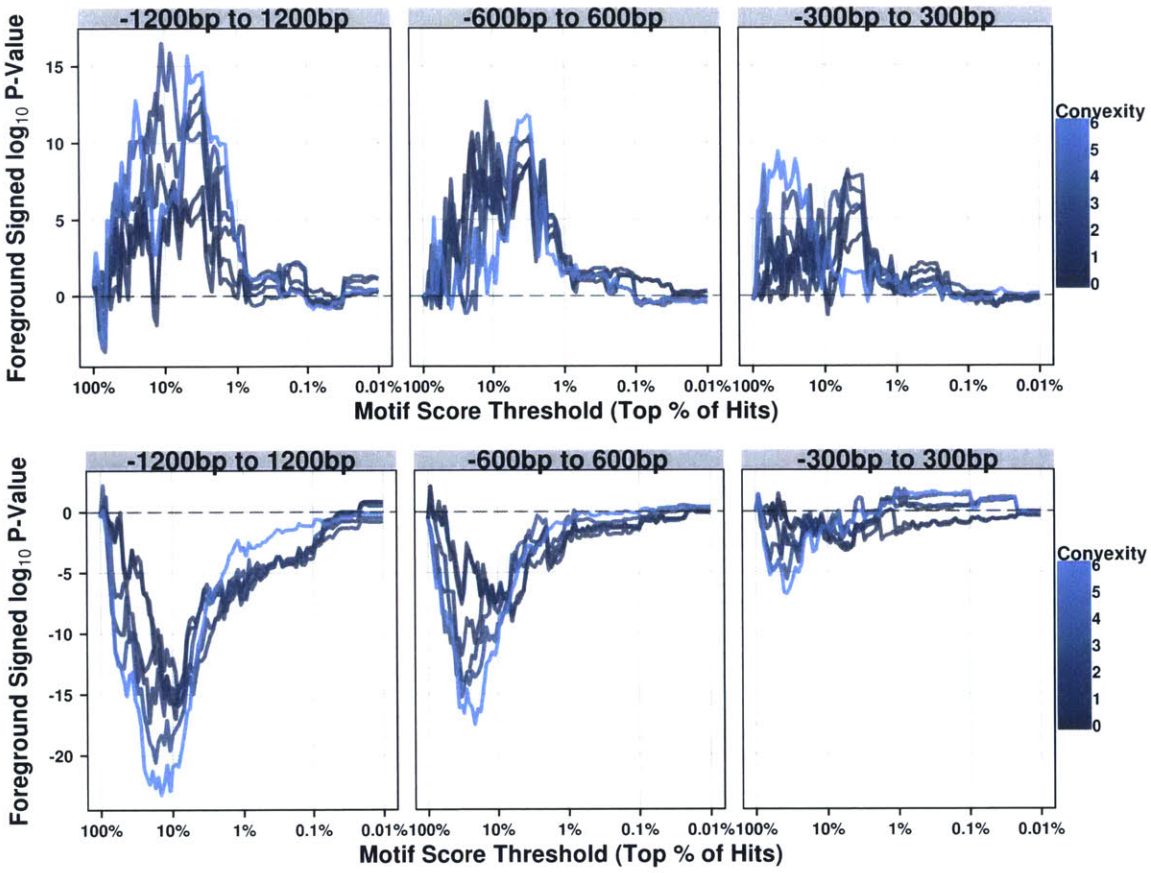

Figure 2-27: First-order expression effects for a range of model score cutoffs across a variety of window filters and profile function convexities for the ATF4-1 and CEBPB motifs.

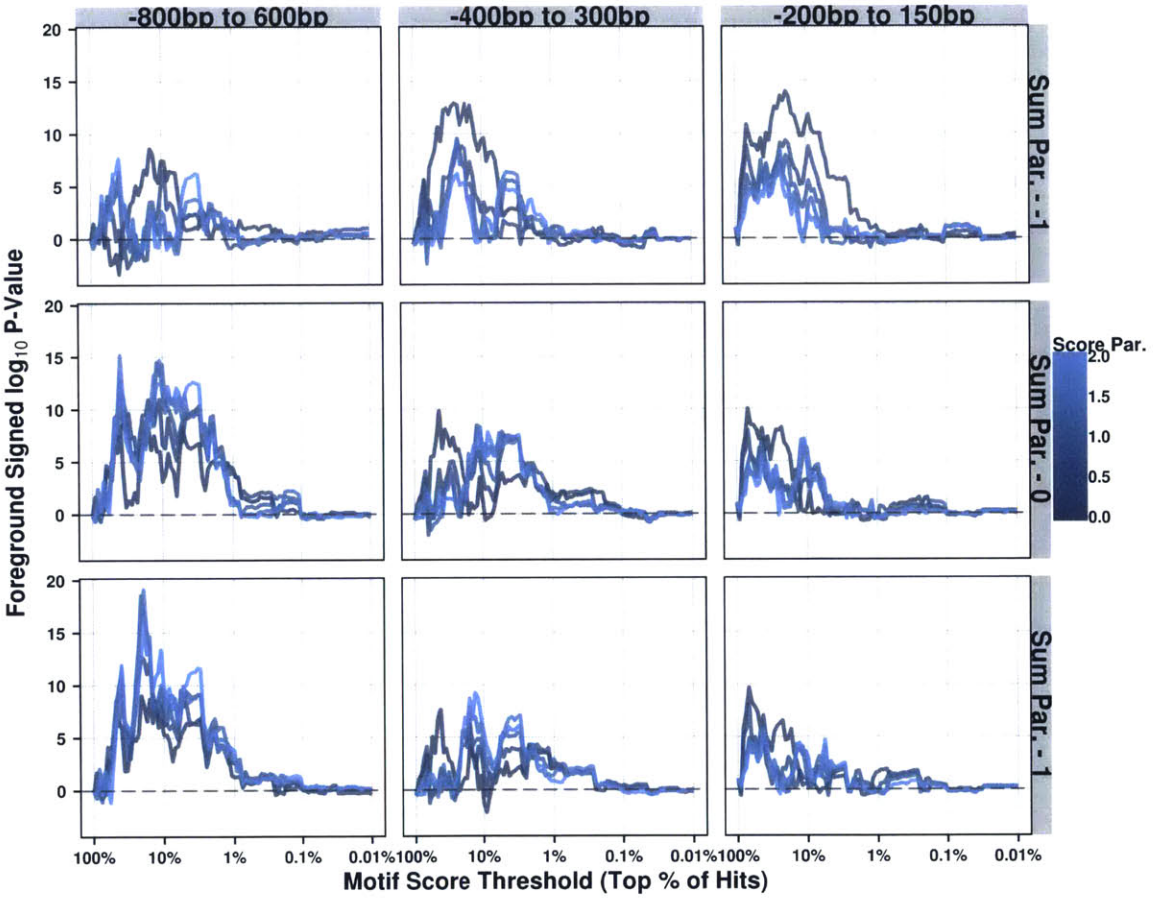

Figure 2-28: First-order expression effects for a range of model score cutoffs across a variety of window filters, sum, and score parameters for the ATF4-1 motif, using a convexity value of 1 corresponding to linear deprecation by distance to TSSs.

range of TSS window filters, suggesting that the score of each motif hit and the number of motif hits do not matter as much as the mean quality of motif hits. Meanwhile, genes regulated by CTCF are not significantly better partitioned by this family of functions, leading us to conclude that the presence model may not be a suboptimal alternative for this factor.

Taken together, these finding further underline the need to fit unique motif hit profile functions for each motif, rather than assuming the same model for each factor as is often done. Each motif associated with active transcription factor exhibits unique profile characteristics in the proximity of genes that are significantly down- or up-regulated. However, we cannot always explain the regulatory behaviour of a given factor using the set of functions we have proposed, which leads us to propose an unsupervised method of constructing profile scoring functions.

2.5 Recovering Regulatory Programs Using Motif Profile Clusters

So far in this chapter we have only considered a binary classification of motif hit profiles in the proximity of transcription start sites. Each time we use a motif hit profile function to score genes for regulatory potential by the transcription factor associated with the motif, we assume that prior distribution underpinning this function is shared by all genes regulated by the TF. Of course, profile functions are fairly flexible in this regard: presence models give a high score to any profile that has at least one motif within a set distance of the TSS, sum models give a high score to any profile that has many low-scoring motifs within this TSS window or a few high-scoring hits, and so on.

However, it is possible that several configurations of motif hits around target genes' TSSs allow for *trans*-regulation by a given TF. Furthermore, we can only test so many different profile functions, while the number of possible such functions is

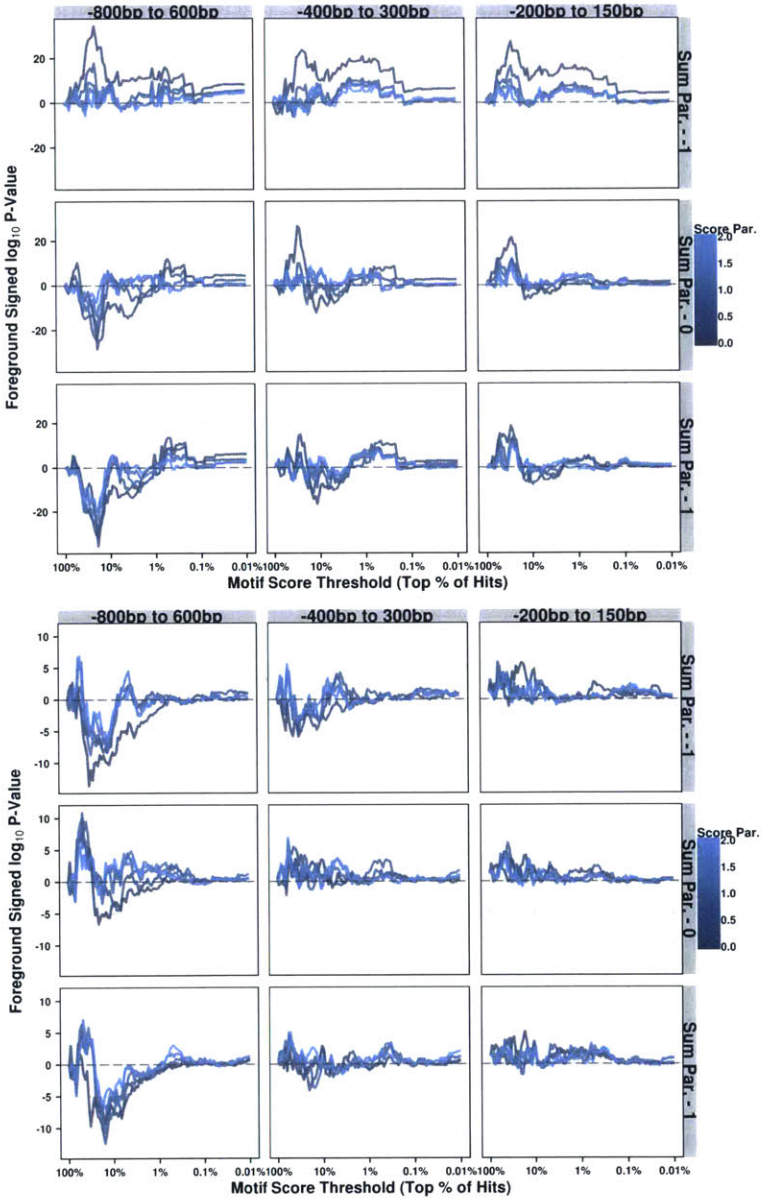

Figure 2-29: First-order expression effects for a range of model score cutoffs across a variety of window filters, sum, and score parameters for the BHLHE40-known3 (top) and CTCF-known1 (bottom) motifs, using a convexity value of 1.

infinite - we have not yet considered functions not centered around TSSs, to give one example. We thus consider an alternative unsupervised method for finding the motif hit profiles associated with regulation by a given TF that relies on hierarchical clustering.

In order to cluster motif hit profiles, we must first define a distance metric between any two profiles. To do this, we first subdivide a wide region around TSSs into smaller overlapping bins. We then calculate the presence of the motif within each bin for each TSS using a position-agnostic metric such as the sum of the motif hit scores occuring in the bin. For each TSS, we have thus obtained a vector representing the presence of the motif in a series of proximal regions. This vector is then normalized relative to background nucleotide bias by calculating the same bin scores for each of the given motif's shuffled versions, taking the bin-wise mean across all shuffled bin vectors, and subtracting this shuffled mean vector from the original motif's bin vector.

By taking, for instance, the Euclidean distance betweem the vectors corresponding to a pair of TSSs we hence obtain a measure of the difference between their motif hit profiles. Calculating this distance between all possible pairs of TSSs, we can apply any one of a number of algorihms that perform agglomerative clustering to obtain groups of TSSs that share similar hit profiles. In our case, we use the complete linkage method as implemented in the `hclust` function included in the `stats` R package (v3.0.1).

This approach builds upon a series of motif clustering analyses previously described in the literature. For instance, Lifanov et al. (2003) clustered motif hits within the bodies of several genes in house flies by counting the number of motif hits occuring in sliding windows, varying motif score thresholds and window sizes to find motif clusters that exhibited the best correlation with experimentally verified *cis*-regulatory regions. Warner et al. (2008) calculated the enrichment of motifs within CREs identified by finding clustering motif occurances to identify the targets of transcription factors within modules of genes linked by GO annotation and coexpression

patterns. Several algorithms have been developed to identify CREs based on clusters of motif hits, such as Cluster-Buster, which takes background nucleotide bias into account by considering the probability of higher-scoring hits associated with suffixes of each motif co-ocurring with the original motif (Frith et al., 2003). This algorithm was recently applied to link TFs to target genes by Janky et al. (2014), who incorporated motif clusters alongside cross-species conservation data to produces ranked lists of genes based on regulatory potential.

Our clustering model improves upon these predecessors in several important aspects. Instead of scanning the genomic areas proximal to the TSSs of a small subset of genes grouped together *a priori* using secondary datasets, we scan the neighbourhoods of all genes' TSSs to find novel modules of genes, and then show that these modules can be linked to differential patterns of expression. Correction for background nucleotide frequency is done using the shuffled counterparts of motifs, which unlike the corresponding method used by Cluster-Buster takes into account the background bias in the region around motif hits, not just directly proximal to them. Finally, by using the sums motif hit scores rather than an arbitrary threshold for motif hits our method is more robust in using the information available through the presence of weak hits.

We performed a hierarchical clustering of the motif hit profiles of ATF4 within 2000kb of TSSs using tiled windows of width 200 base pairs placed at intervals of 5 base pairs. Although most genes had null hit profiles, there emerged clear groupings of genes with peaks of ATF4 motif hits placed at similar distance from their TSSs (Figure 2-30). Furthermore, several of these clusters are associated with elevated expression levels, indicating that they correspond to biologically relevant patterns of regulation.

For instance, there are two clusters corresponding to peaks located 500 base pairs upstream of TSSs and 250 base pairs downstream of TSSs, containing 104 and 68 genes respectively that how significantly higher expression than the genes in the "null" cluster. Likewise, there is a cluster containing 951 genes corresponding to a

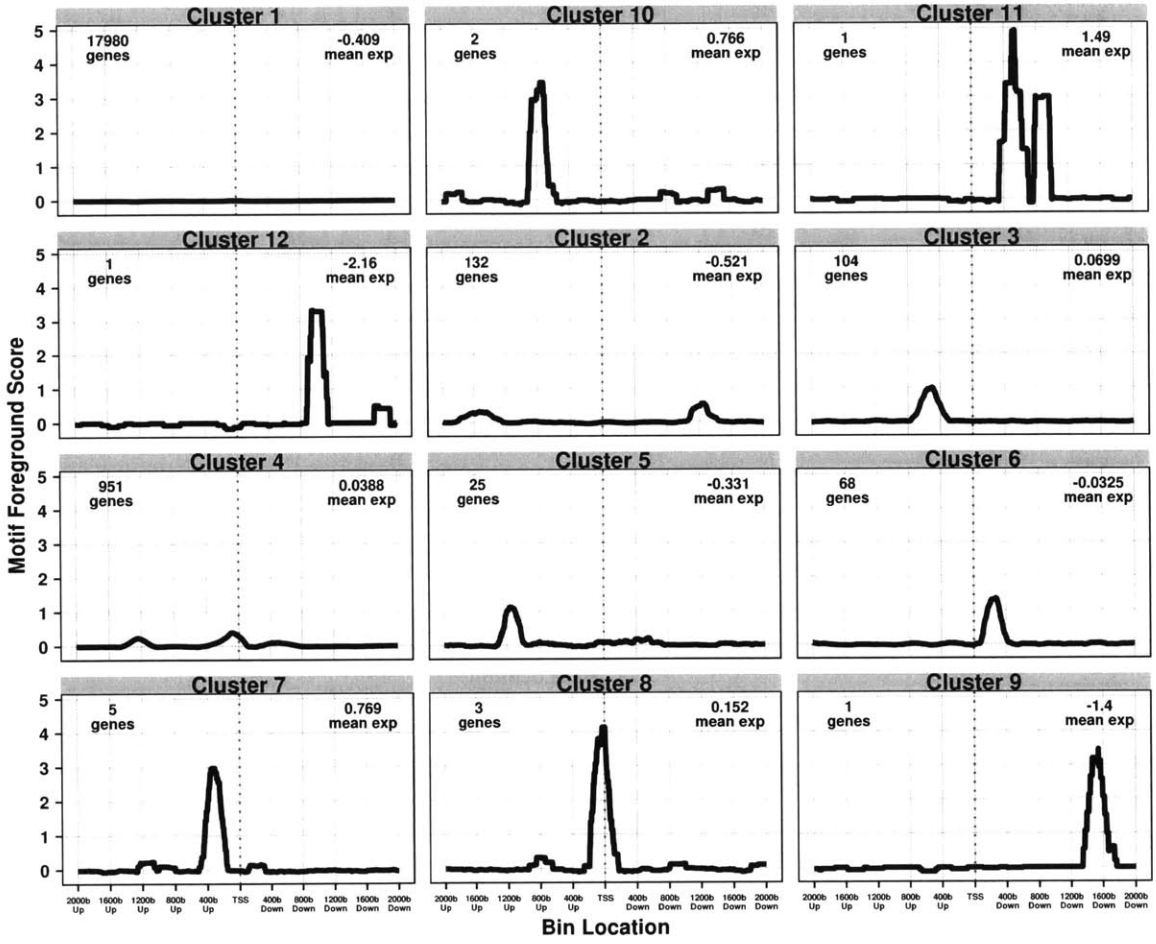

Figure 2-30: The background-normalized motif hit profile clusters generated for ATF4-1 using a window filter of 2000 base pairs around TSSs, with measurements taken in sliding windows of width 200 base pairs located 5 base pairs apart.

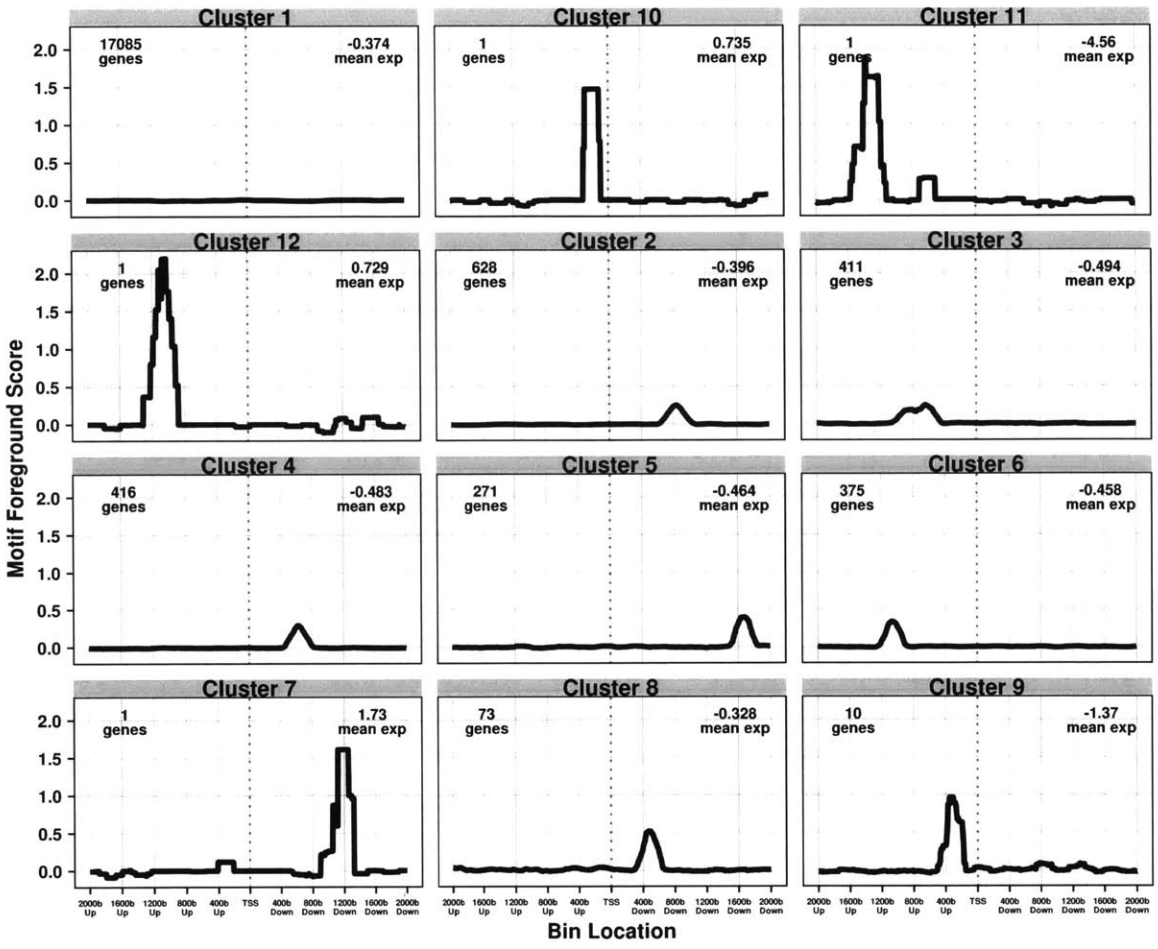

Figure 2-31: Motif hit profile clusters generated for ATF4-2 using the same measurements as above.

bimodal pattern of foreground ATF4-1 motif presence, with peaks 100 and 1250 base pairs upstream of TSSs, that is also associated with activated expression.

The ATF4-2 motif also displays evidence of motif clustering in the proximity of TSSs; however, these clusters are not associated with differential patterns of expression (Figure 2-31). It is thus apparent that motif clustering by itself is not a sufficient condition for regulatory activity. Clusters may need to overlap with CREs in order to influence expression, which suggests a major flaw in the approaches listed above which rely solely on clustering of motifs to identify CREs in the first place.

Clustering of several other key transcription factors in lung tissue reveals a number of other clusters associated with activation or repression relative to the background cluster (Figure 2-32). An important feature of these clusters is that their positions, shapes, modality, and effect on expression is unique to each TF and motif, which further underlines the necessity of fitting motif hit profile functions that are factor-specific. Some transcription factors, such as ELF3, are associated with motifs that impact expression through subtle patters of enrichment in which hits are spread out over multiple peaks of relative small magnitude in the proximity of TSSs. Others, such as ATF4 and BHLHE40, have motifs that have sharp, well-defined unimodal patterns of enrichment that are usually (but not always) located at positions overlapping TSS. The discrepancy between the clusters for the two known CTCF motifs, which are similar to one another, suggests that they may be linked to different binding patterns for the zinc finger protein that fulfill different regulatory roles.

This method of clustering motif hits using sliding windows with local background nucleotide bias correction using shuffled motifs has thus been proven to be able to identify meaningful groups of target genes for transcription factors, as validated by expression profiles. It serves as a useful counterpart for identifying patterns of binding that would be difficult to uncover by applying standard assumptions about the properties of *trans*-regulation with regards to proximity to TSSs, spacing of motif hits, and so on. In the next chapter, we will have the opportunity to test whether the peaks of motif hit occurance we identified using clustering of profiles are concordant with the effect of independently identified *cis*-regulatory elements on expression.

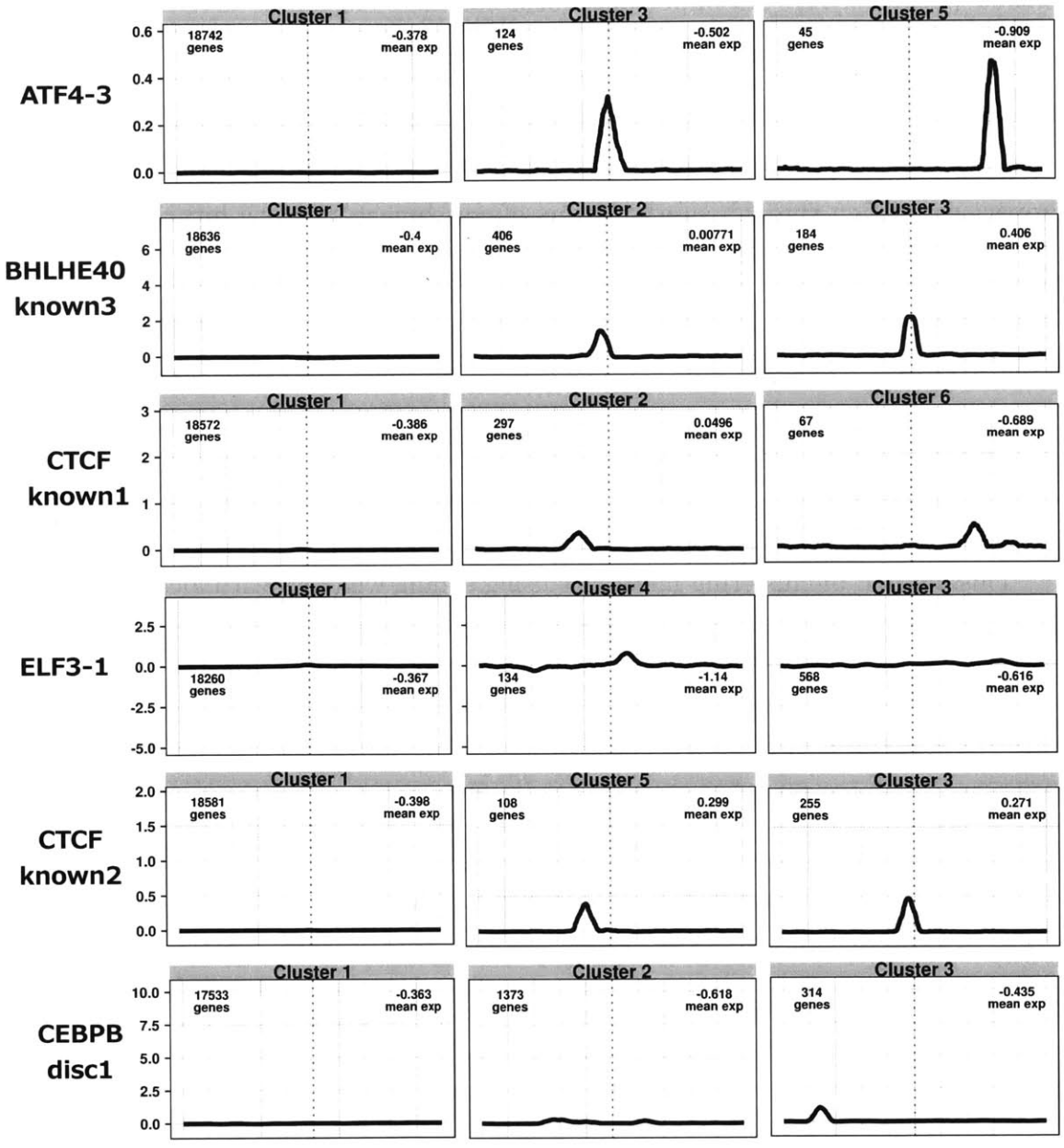

Figure 2-32: Selected motif hit profile clusters generated for various motifs.

Chapter 3

Incorporating Regulatory Regions into Models of *Trans*-Regulation

We have thus far discussed the expression effects of motif hits occuring in the proximity of transcription start sites without considering the presence of *cis*-regulatory elements. Given that *trans*-regulation and *cis*-regulation go hand in hand with one another in controlling the transcription of target genes, any complete model of regulation must consider the effect of both CREs and TF binding, as well as the interaction effects between the two. Having constructed frameworks for robustly measuring transcription factor occupancy proximal to TSSs using motif hits, we now augment them by including regulatory region calls.

Previously described models of regulation display just as much variation in how they incorporate the overlap between putative transcription factor binding sites and regulatory regions as they do in measuring the presence of binding sites. They range from models which do not attempt to consider the interaction of TF activity with CREs (Glass et al., 2015), models which use DNase1 footprints and TSS proximity as a proxy for *cis*-regulatory potential (Neph et al., 2012), models which use DNase1, TSS proximity, conservation scores, and repressing histone marks as a proxy for *cis*-regulatory potential (Pique-Regi et al., 2011), and models which use a variety of

histone marks in isolation (Cheng et al., 2012). The regulatory regions we use are based upon an integration of DNase1 data and various histone marks, and hence leverage the interplay between several of the features that are only incorporated in isolation within the models listed above. Furthermore, we have already introduced TSS distance as a explanatory variable in our presence models, instead of treating it as a static cutoff for *cis*-regulatory potential as in several of these models.

3.1 Effect of *Cis*-Regulatory Presence on Expression Profiles

Before considering expression effects associated with the predicted binding of individual transcription factors, we first examine what effect the mere presence of *cis*-regulatory elements has on expression profiles. We use the elements predicted by the HoneyBadger2 algorithm, which uses DNase1 footprints to delineate the genome into regions that are then annotated using a Hidden Markov Model based upon a set of histone mark calls (Appendix A.3). These predictions are specific to 127 tissues associated with the Roadmap and ENCODE epigenetic consortia, a number of which are related to those from which our expression datasets were drawn.

Histone modification marks and DNase1 data have repeatedly been proven useful in predicting *trans*-regulation. For example, Ernst and Kellis (2013) demonstrated the existence of signature chromatin states for individual transcription factors that encoded a preference for binding above and beyond the presence of sequence motifs associated with the TFs. (O'Connor and Bailey, 2014) used the correlation between histone marks and expression at proximal genes to link identify genes which were targets of CREs overlapping with the marks. DNase1 footprints have been used improve the quality of transcription factor binding site predictions based on sequence motifs (Sherman and Cohen, 2012). Because the regulatory regions we use integrate histone marks and DNase1 footprints, their presence should have tissue-specific effects

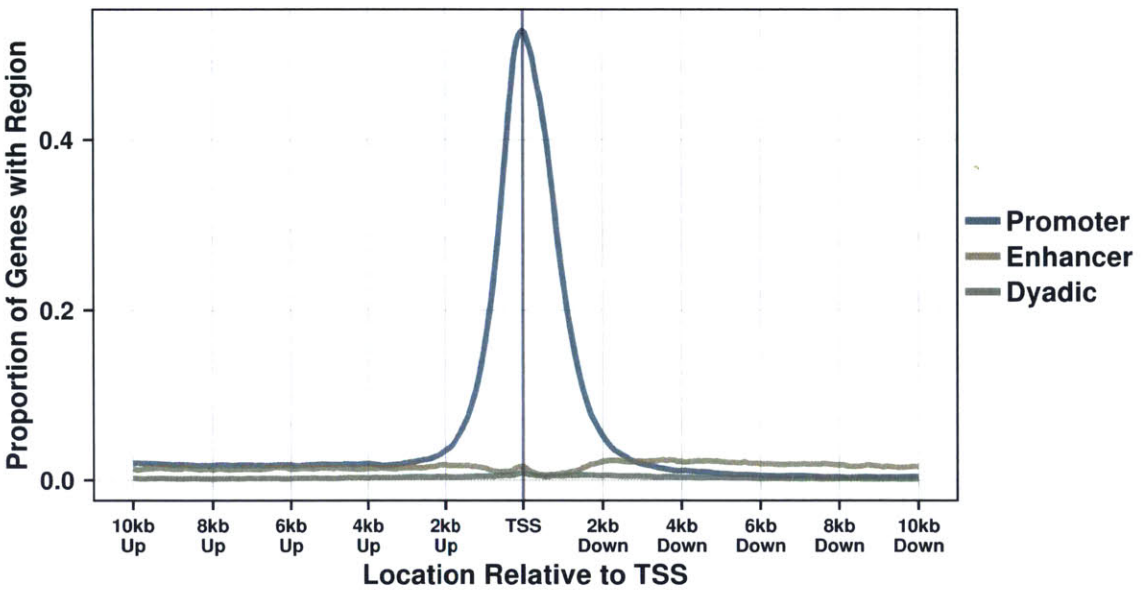

Figure 3-1: The proportion of transcripts with each type of regulatory region present a certain distance upstream or downstream of the transcription start site.

on both expression and coexpression with potential regulators.

3.1.1 First-order Expression Effects of Regulatory Regions

We first examine the expression effects of regulatory elements predicted for lung tissue (Roadmap tissue E096) in the proximity of gene transcription start sites. As one would expect, promoters are enriched at transcription start sites, while the presence of enhancers and dyadic regions follows a flatter pattern (Figure 3-1). Surprisingly, the peak of promoters at TSSs is symmetric, and not centred upstream of TSSs, indicating that many promoters overlap with the body of the gene.

We thus check whether promoters (and other regulatory regions) that overlap with gene bodies tend to be enriched in areas of genes that are not transcribed. As shown in Figure 3-2, this is not the case: promoters occur more frequently in coding regions of genes than in introns, although the highest frequency of promoters is observed in untranscribed portions of exons. This effect is persistent across the entire span of

transcript bodies from transcription start sites to the 3' end of the final exon of each transcript, but promoters become steadily less likely to occur towards 3' end of the transcript. On the other hand, enhancers are roughly half as likely to occur in coding regions compared to introns and UTRs in the 5' half of transcript bodies, and equally as likely to occur among the different gene elements in the 3' end, while dyadic regions follow an intermediate pattern.

Although we observe that *cis*-regulatory elements have certain patters of enrichment relative to transcription start sites and gene structural elements, this does not necessarily mean that these elements are active in regulating transcription in lung tissue. We thus measure the effect that the presence of CREs has on patterns of expression by themselves (Figure 3-3). Not only are promoters enriched in presence in the area just upstream of transcript starts, their presence in this area has a strong effect on expression levels, as one would expect. However, within the 5' half of transcript bodies, although promoters occur less frequently in introns, their presence within introns has a stronger activating effect on expression than in exons. In the 3' half of transcript bodies, overlap between introns and promoters represses transcription, as does overlap between exons and promoters. Promoters in untranscribed portions of exons also have an activating effect on expression that does not become repressive in the latter half of transcript bodies.

The presence of enhancers that overlap with introns has a relatively constant activating effect on expression throughout transcript bodies that is only slightly stronger than their effect outside of transcript bodies. This effect is wiped out in the area roughly 1000 base pairs around TSSs, possibly due to the competing presence of promoters. Enhancers overlapping with exons are relatively rare and thus their effect on expression is difficuly to measure; however, it appears this effect is nugatory in the 5' half of transcript bodies and and mildly activating in the 3' half.

Finally, the observed effect of dyadic regions on transcript expression is much more unstable than that of promoters and enhancers with respect to distance from

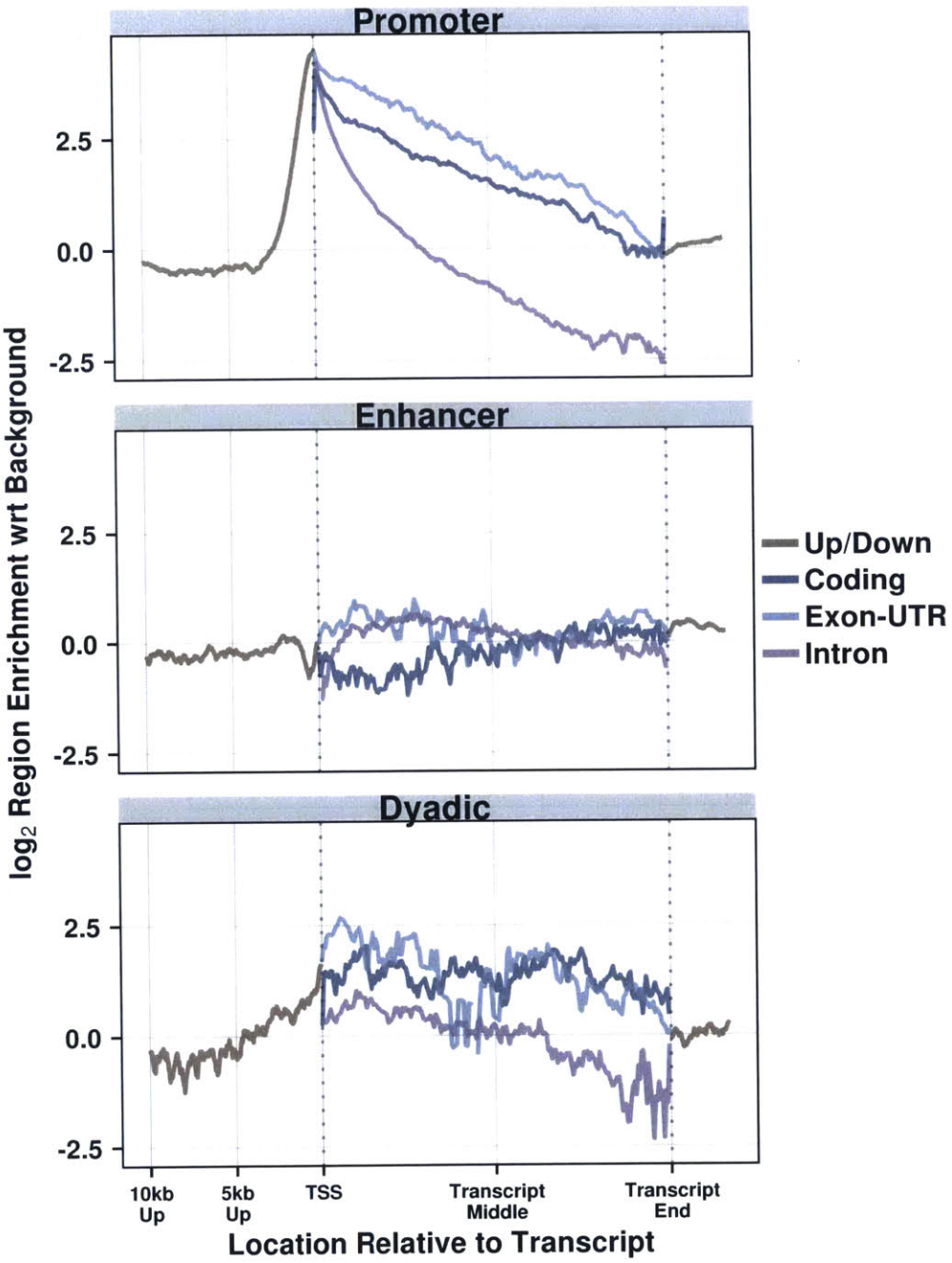

Figure 3-2: Enrichment of regulatory regions in the proximity of the TSS according to overlap with gene structure elements. Position outside of transcript bodies is measured in absolute base pair terms, while position within trasnscript bodies is measured relative to the position relative to TSS and the end of the final exon of each transcript. Enrichment is calculated relative to the mean of region presence 10kb upstream and 2kb downstream of TSSs.

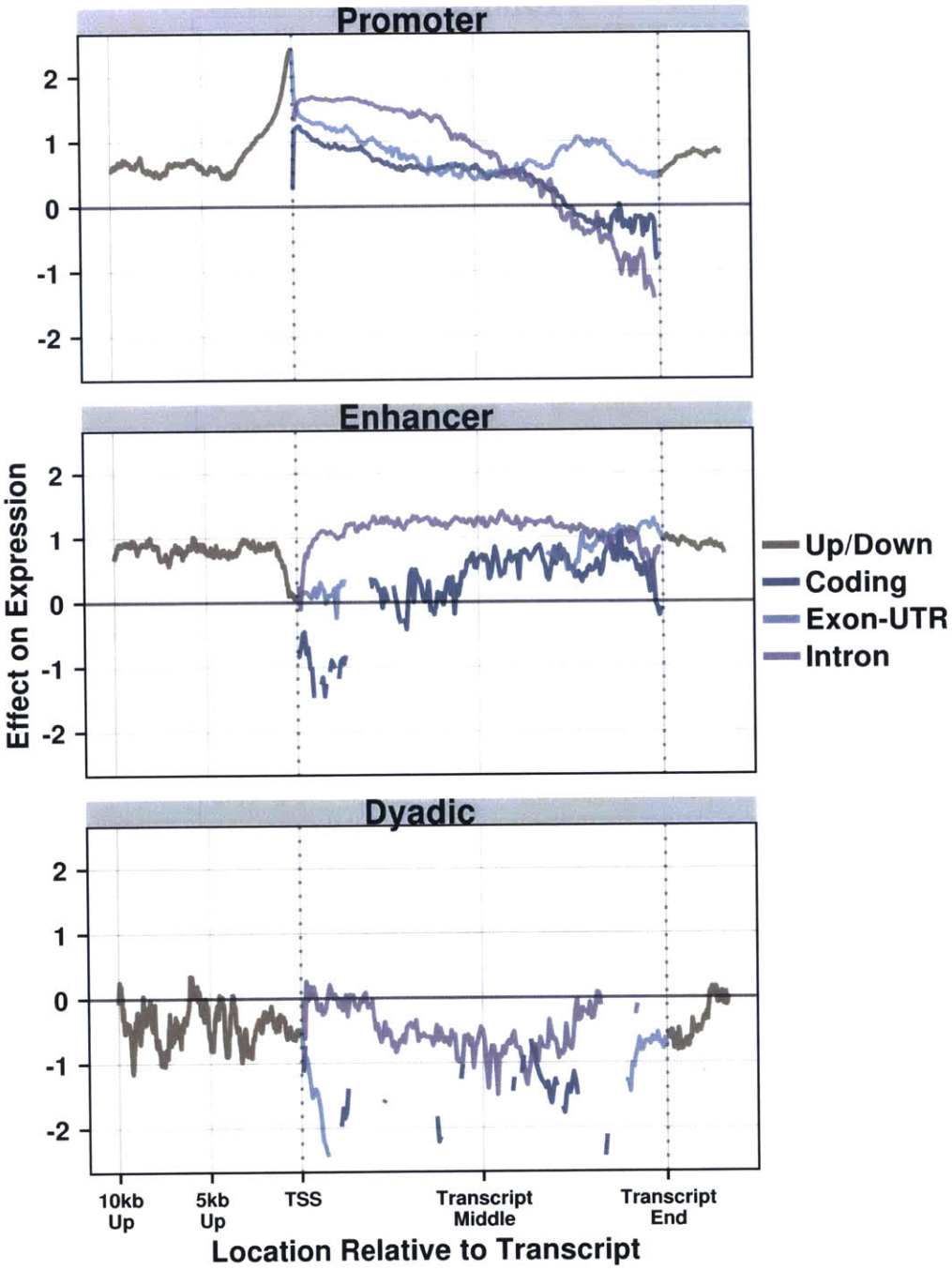

Figure 3-3: Mean expression of gene transcripts partitioned according the absence or presence of the three types of regulatory regions overlapping structural elements at a certain distance upstream or downstream of their transcription start site. Note that locations where fewer than 0.1% of transcripts had a regulatory region present were omitted.

transcription start site. This is most likely at least partly due to the fact that these regions are less likely to occur in general, and so their effect is measured over a smaller number of transcripts and hence is more vulnerable to random noise. Nevertheless, dyadic regions appear to have a repressive effect on expression, except for a region extending about a quarter of the way along each transcript body starting at the TSS, when the effect is close to zero.

It is therefore apparent that each type of regulatory region has a unique set of effects on expression of proximal genes based on their position relative to the transcript, independent of the binding of any particular *trans*-acting factors. These effects are modulated by overlap with structural elements within transcript bodies, with regulatory regions occurring within introns having a stronger effect on expression due to their lack of interference with regions that are involved in coding for proteins.

3.1.2 Second-order Expression Effects of Regulatory Regions

We have thus established a baseline for the effect of *cis*-regulatory elements on expression levels independent of *trans*-regulatory events. We can likewise establish such a baseline for second-order expression effects related to individual transcription factors. In particular, we consider coexpression with the set of transcription factors that are both highly expressed in the lung dataset and for which motifs are available to be able to pinpoint potential binding sites. These factors are most likely to play a regulatory role in lung tissue, and moreover, we will be able to measure the expression effects of binding sites associated with these TFs identified using motifs in subsequent analyses.

For example, ATF4, the most highly expressed such factor in the lung dataset, exhibits several clear relationships between its coexpression profiles with potential targets and the presence of regulatory regions in the proximity of these targets (3-4). As expected, promoters appearing just upstream of transcription start sites mediate coexpression with ATF4, and enhancers appearing in the same location decrease coexpression. However, promoters overlapping with introns within the 5' half of target

transcript bodies have an even stronger effect coexpression. Enhancers coocurring with introns also lead to higher coexpression with ATF4 than background levels when they are placed in the middle of transcript bodies.

The second-order expression effects of regulatory regions that overlap with other structural elements are more difficult to elucidate. In the case of promoters, overlap with exon UTRs leads to higher coexpression in the 5' third of transcript bodies, lower coexpression in the middle third of transcript bodies, and then higher coexpression again in the 3' third. Overlap with coding portions of exons leads to unstable second-order effects that roughly correspond to background levels, and can hence probably dismissed as noise. Overlap between enhancers and exons is sufficiently rare that observed second-order expression effects in this case are also overwhelmed by noise.

We further analyze whether these second-order expression effects are sensitive to whether we exclude genes that are not highly expressed. In the above analysis, we included all protein-coding genes located on the non-sex chromosomes. When we exclude genes that according to mean expression across all samples, we see that co-expression effects are indeed affected by the noise inherent in measuring coexpression with genes that have low expression (Figure 3-5). Second-order expression effects of promoters overlapping with introns become more muted relative to the peak observed at TSSs, and the effect of promoters overlapping with coding regions converges to zero in the 3' half of transcript bodies. This is concordant with how we would expect promoters to behave when modulating the *trans*-regulation of target genes by ATF4; we will thus use only the top half of genes by mean expression when calculating second-order effects in order to avoid the confounding effects of lowly-expressed genes. The effect of mean expression thresholds on second-order expression effects of enhancers within transcripts are harder to measure given the rarity of enhancers that co-occur with coding regions; however, enhancer that overlap with introns also seem to have a slightly lower effect on coexpression with ATF4 for highly-expressed genes.

In conjunction with our findings so far for first-order expression effects, these

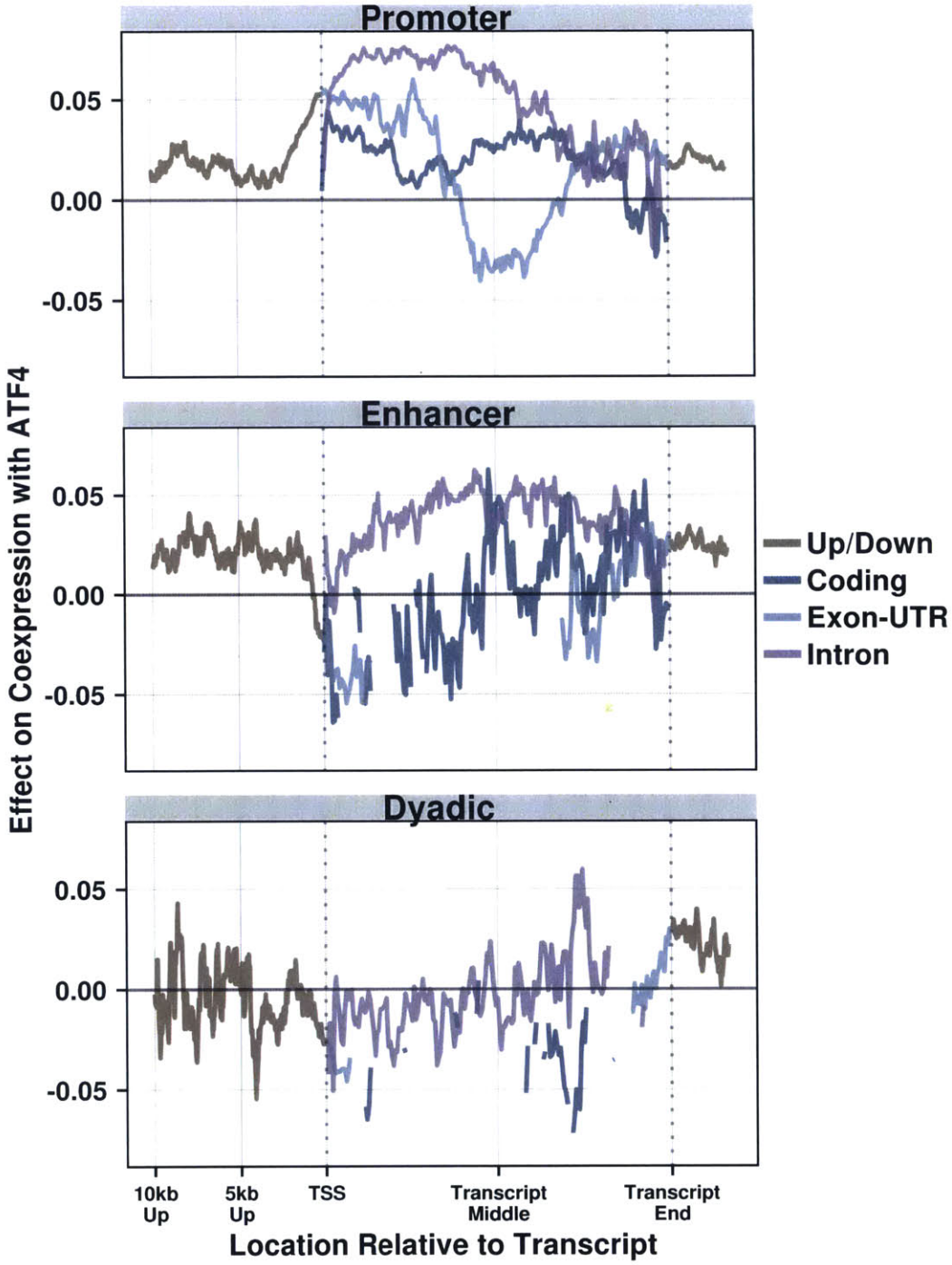

Figure 3-4: Mean coexpression with ATF4 of transcripts partitioned according the absence or presence of regulatory regions with respect to distance relative to TSS and overlap with transcript structural elements.

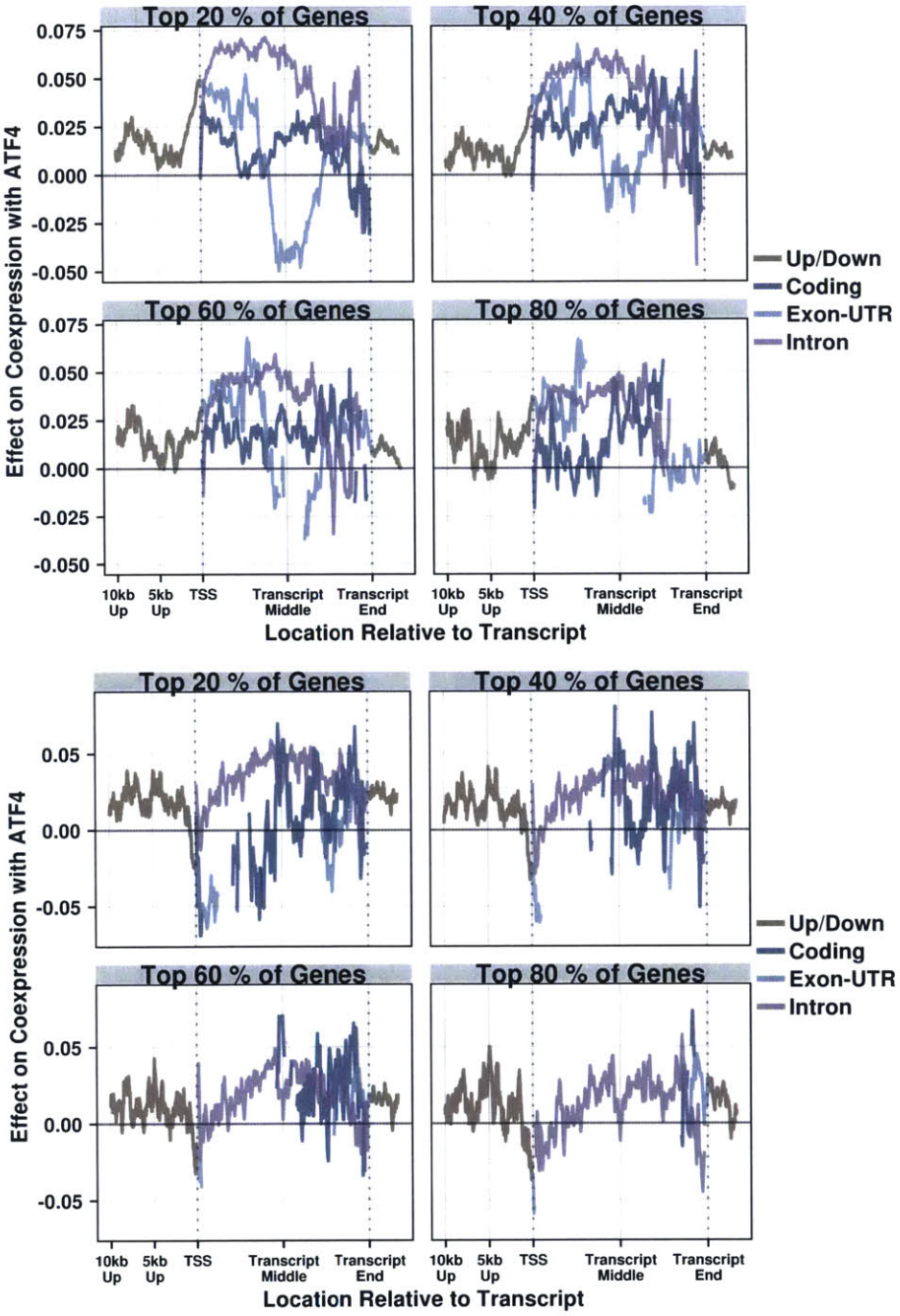

Figure 3-5: Mean coexpression with ATF4 of transcripts partitioned according the presence of promoters with respect to transcript structure, for subsets of target genes according to mean expression cutoffs.

patterns in second-order expression effects suggest that *cis*-regulator presence within transcript bodies plays a significant role in *trans*-regulatory circuits. Although the canonical location of promoters just upstream of transcription start sites does have a clear relationship with TF coexpression, promoters which appear downstream of TSSs must also be taken into account. As we have already seen, promoter presence is distributed more or less symmetrically about TSSs, and expression effects suggest that although relative rare, promoters that are located deep within transcript bodies have a clear effect.

However, when we examine the same relationship between regulatory region presence and coexpression with other TFs that are highly expressed in the lung tissue dataset, we see that these observations are not necessarily true for TFs in general (Figure 3-6). Indeed, transcription factors exhibit various combinations of second-order expression effects with respect to distance from TSS and the overlap between regulatory elements and coding regions. Transcription factors such as HERPUD1 have high coexpression with transcripts that have promoters located just upstream of their TSS, but also with transcripts that have promoters located within transcript bodies depending on overlap with introns and exons. Other TFs such as BHLHE40 have second-order expression effects that "flip" in the middle of transcript bodies, with promoters in the 3' end increasing coexpression rather than decreasing it; based on the decrease in coexpression with BHLHE40 seen in genes that have proximal upstream promoters, we can infer that BHLHE40 acts as a repressor in lung tissue and that promoters in the 3' end of transcript bodies cancel out this effect.

Meanwhile, CEBPB and CEBPD both behave similarly, as one would expect based on their shared role as CCAAT/enhancer-binding proteins. Coexpression with both of these TFs in increased in transcripts with enhancers within their bodies, an effect not present in either of BHLHE40 or HERPUD1. Furthermore, second-order expression effects for both of these TFs is greatly increased in the presence of promoters within transcript bodies compared to the presence of promoters outside of transcript bodies.

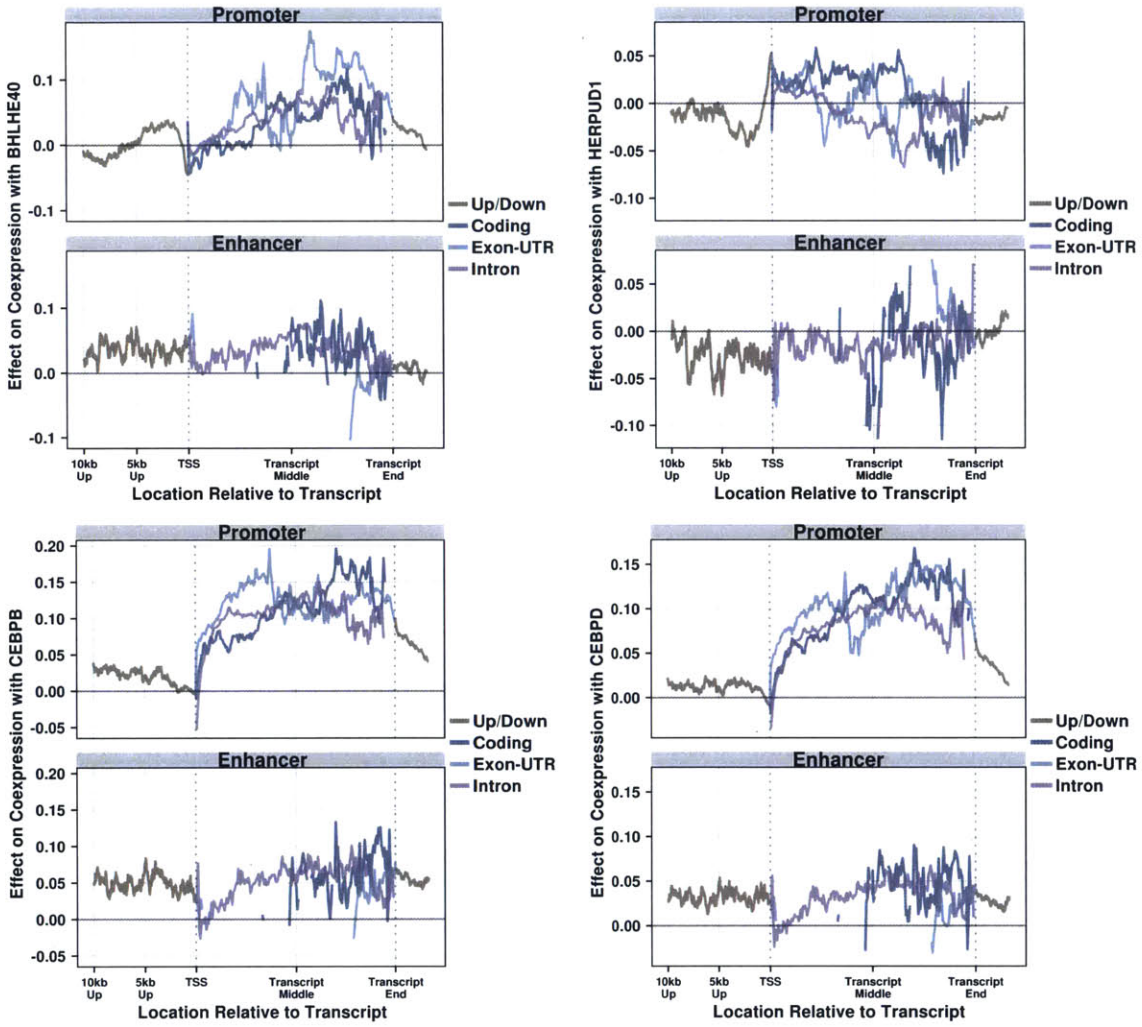

Figure 3-6: Effect of regulatory region presence on the coexpression of highly-expressed genes with a selection of transcription factors highly expressed in lung tissue according to structural area overlap.

Another way to check the relevance of the coexpression effects observed for ATF4 is to compare them with the corresponding effects for genes that are not transcription factors, but which have a similar mean and variance of expression among the lung tissue samples(Figure 3-7). We thus observe a background "null" of coexpression effects associated with genes which should not have any direct interaction with CREs that can be attributed to larger networks of coexpression present in lung tissue. For these transcripts, pseudo-second-order effects are also high within transcript bodies that contain promoters, suggesting that genes with this property may tend to be coexpressed in general due to relationships not involving *trans*-regulation.

Given that these profiles closely resemble that of ATF4, CEBPB, and CEBPD, we can call into question whether these three factors have a significant regulatory role in lung tissue, or are simply coexpressed with target genes indirectly through association with other active factors. On the other hand, the coexpression profiles of HERPUD1 and BHLHE40 point towards several hallmarks of genuine regulatory activity which include a sharp peak of coexpression effect of promoters around the TSS, declining and generally low coexpression effects along the transcript body, and insignificant coexpression effects for enhancers in general within and around transcript bodies.

Without using any information unique to individual transcription factors, we have nevertheless been able to observe the binding properties of several TFs in lung tissue. Even for TFs other than ATF4, which have coexpression effects concordant with regulatory activity, the distances over which these effects are significant varies between factors. We now introduce a way of gauging whether or not these effects are actually brought about by binding using motifs associated with these TFs, which should have some ability to predict specific binding sites. In similar fashion we seek to prove that the coexpression effects observed for ATF4 are unlikely to be associated with actual binding of this factor in a regulatory capacity.

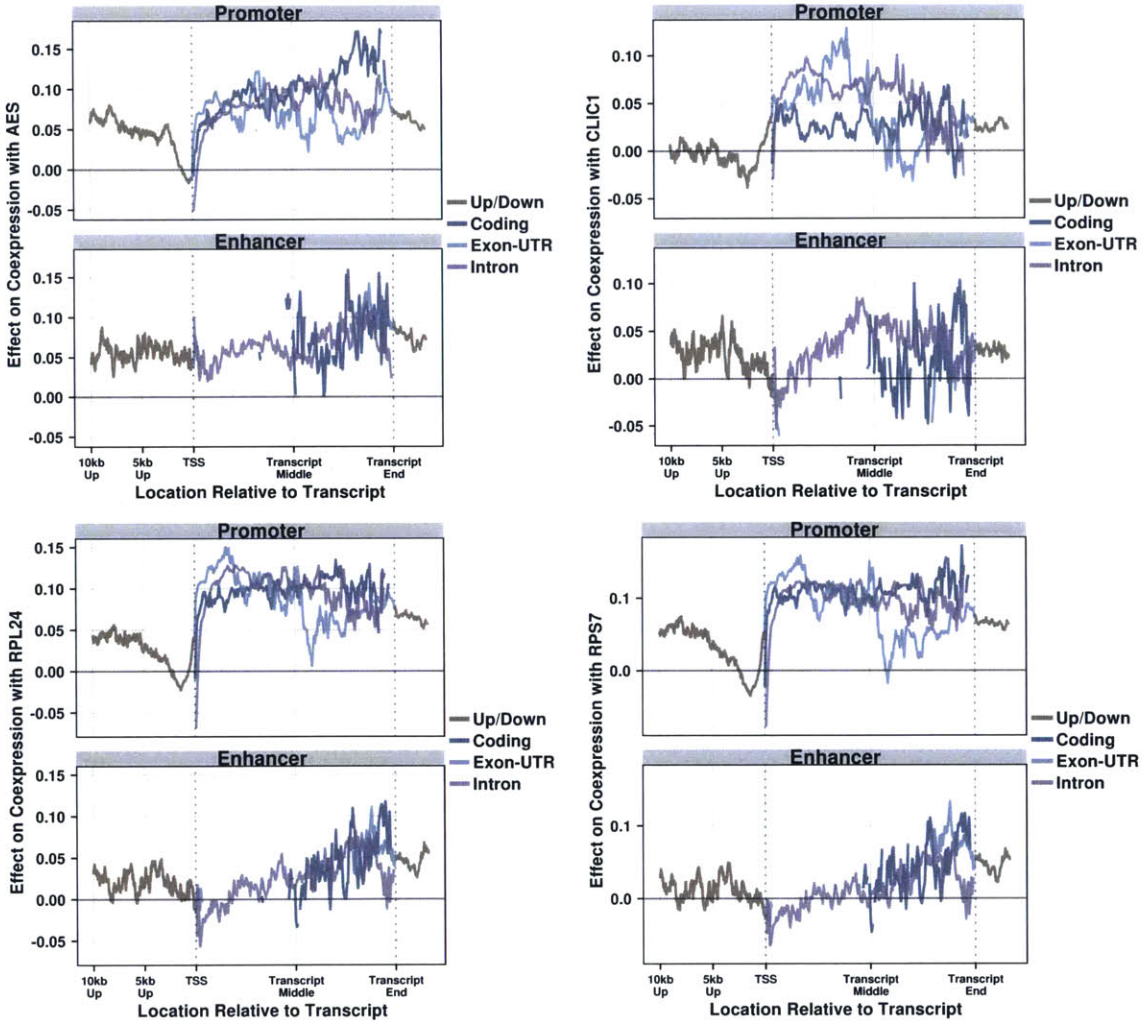

Figure 3-7: Effect of regulatory region presence broken down by structural area overlap on the coexpression of highly-expressed genes with a selection of genes that are not transcription factors but have a similar expression profile to ATF4.

3.2 Measuring the Influence of CREs and Motif Hits on Expression Profiles

We have now tried to observe the expression effects of *trans*-regulation in two orthogonal ways. First, we checked what effects the presence of *cis*-regulatory elements around transcription start sites and within transcript bodies have on overall expression levels and on coexpression with potential transcription factors. We then checked whether the presence of at least one motif associated with a transcription factor in the proximity of transcription start sites have the same effects. In both cases, we were not able to find conclusive evidence that a particular pattern of CREs or motif hits around a transcript's start site has an effect on its expression profile. We now attempt to merge the two approaches, with the hypothesis that *trans*-regulation of target genes depends on the interaction between *trans*-acting proteins, whose binding can be approximated using motif hits, and CREs.

A fundamental difference between the way we measured the expression effects of CREs and motif hits is that for the former, we considered the effect of a CRE occuring exactly at a particular distance from the TSS, while for the latter, we considered the effect of any hits occuring within a particular distance from the TSS. This was done due to the fact that CREs are by nature regions that span sections of the genome that are measured in hundreds if not thousands of base pairs, while motif hits are represented as point events along the genome. We must thus use wide bins to capture the presence of motif hits, as most motifs match sparsely across the genome, as opposed to CREs, whose presence can be expected in a meaningful subset of genes at any particular distance from their TSSs, especially in the case of promoters and enhancers.

There are hence several ways to measure the effect that the overlap between CREs and motif hits has on patterns of expression and coexpression. The simplest approach is to break down motif hits according to overlap with regulatory regions (or lack

thereof), and apply our presence model as before to each set of hits separately. As shown in Figure 3-8, expression effects of the HERPUD1 motifs thus measured seem to be dominated by overlap with regulatory region, as the original and shuffled motifs have the same effect profile for each region. On the other hand, there does appear to be window size of about 800 base pairs for which HERPUD1 motifs that overlap with promoters have a significant effect on coexpression over and beyond that of their shuffled counterparts.

However, given that this effect happens over such a narrow range of window sizes, one is inclined to question whether it is robust, or merely a fluke based on the datasets we are using. Furthermore, what proportion of this effect is attributable to the presence of promoters, and what proportion can be credited to the presence of the HERPUD1 motif within this region? To answer these questions, we construct a framework that can robustly measure the effect that the interplay of regulatory regions and motif hits in a given region have on expression effects.

Given a set window in the vincinity of TSSs, we would like to know what effect on expression effects the presence of motif hits overlapping with a given type of regulatory region have independent of background regulatory region presence, background motif presence, and background nucleotide bias, as measured by the presence of shuffled variants of the motif. We thus use the model

$$E(t) \sim Reg + P(shuf_{noreg}) + P(shuf_{reg}) + P(orig_{noreg}) + P(orig_{reg}) \qquad (3.1)$$

where $E(t)$ is the expression of the proximal transcript, $P(shuf_{noreg})$ and $P(shuf_{reg})$ is the motif hit profile function we are using applied to the shuffled motif variants inside the window not overlapping with the regulatory region and overlapping with the region respectively, and $P(orig_{noreg})$ and $P(orig_{reg})$ is the same function applied to the original motif not overlapping and overlapping with the regulatory region. We use a linear fit to solve for the coefficients of this model for using all of the transcripts

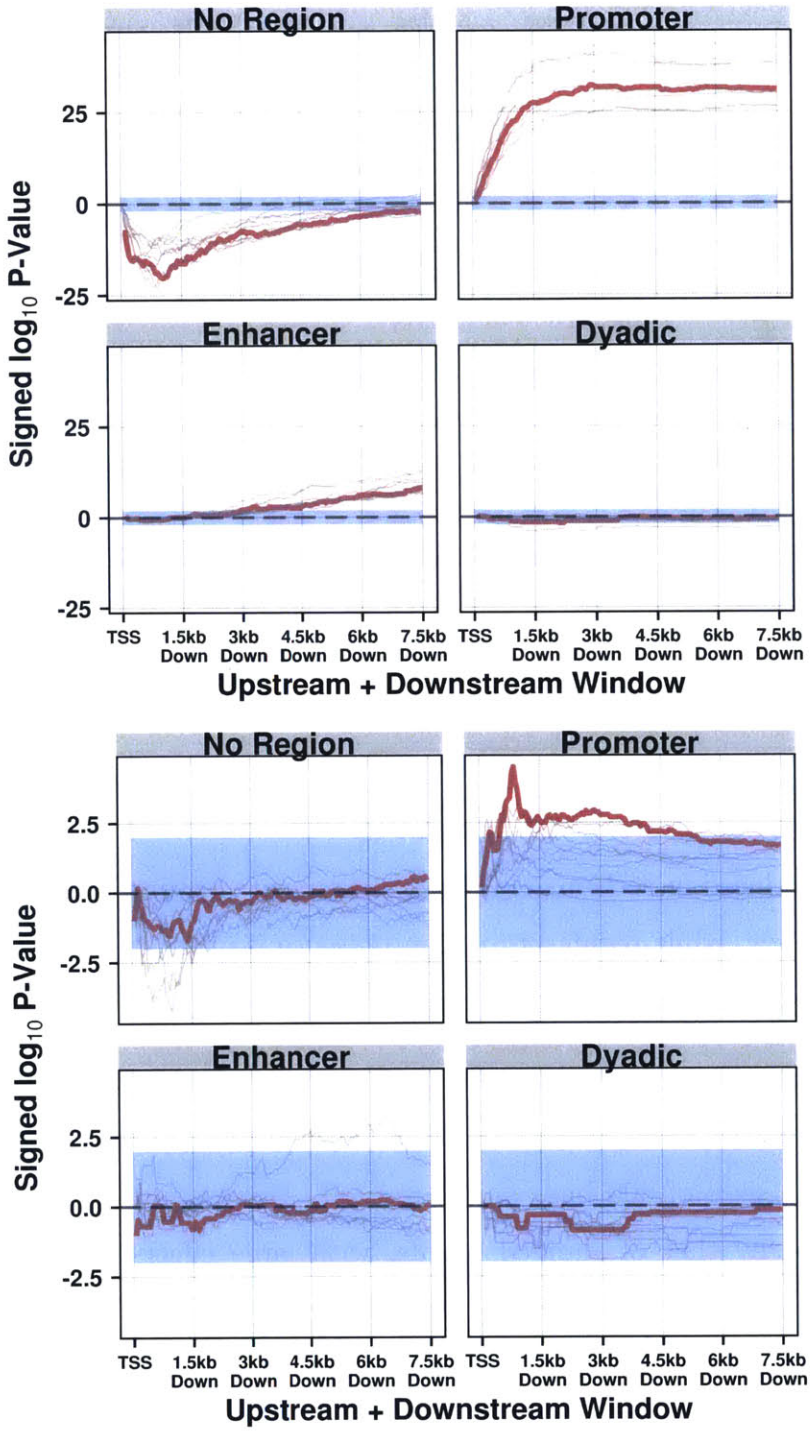

Figure 3-8: First-order (top) and second-order (bottom) expression effects, as measured by signed p-value, for the HERPUD1 motif and its shuffled variants broken down by overlap with regulatory regions.

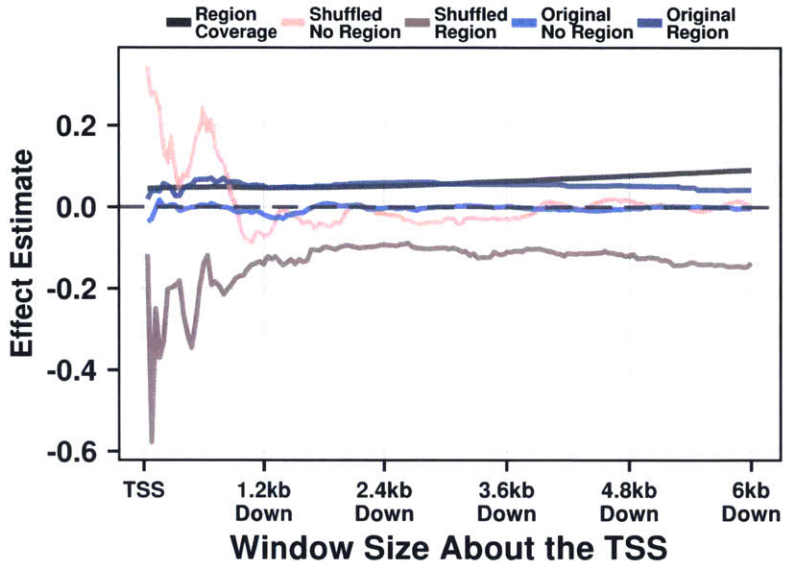

Figure 3-9: Second-order expression effect model coefficients for the HERPUD1 motif for various sizes of windows centred at transcription start sites.

available, with a separate model being fit for each window. The motif hit score sum function introduced in Section 2.4 is used to aggregate hits into a single numeric value for each window, while regulatory region presence is calculated as the proportion of bases in the window located within the given type of region.

As shown in Figure 3-9, there is indeed a peak of the strength of the effect that HERPUD1 motifs overlapping with promoters have on proximal gene coexpression when window sizes of 800 base pairs are used. However, this effect appears to be overshadowed by that of shuffled motifs, with large effects in opposite directions observed for shuffled motifs occuring inside and outside of promoters. We thus check the errors included in these fitted values to test their statistical significance (Figure 3-10). This reveals that although the effects attributed to the presence of regulatory regions are highly robust, those of shuffled motifs are not for any window size, and those corresponding to the original motif are barely significant, but maximally so at window sizes of 800 base pairs.

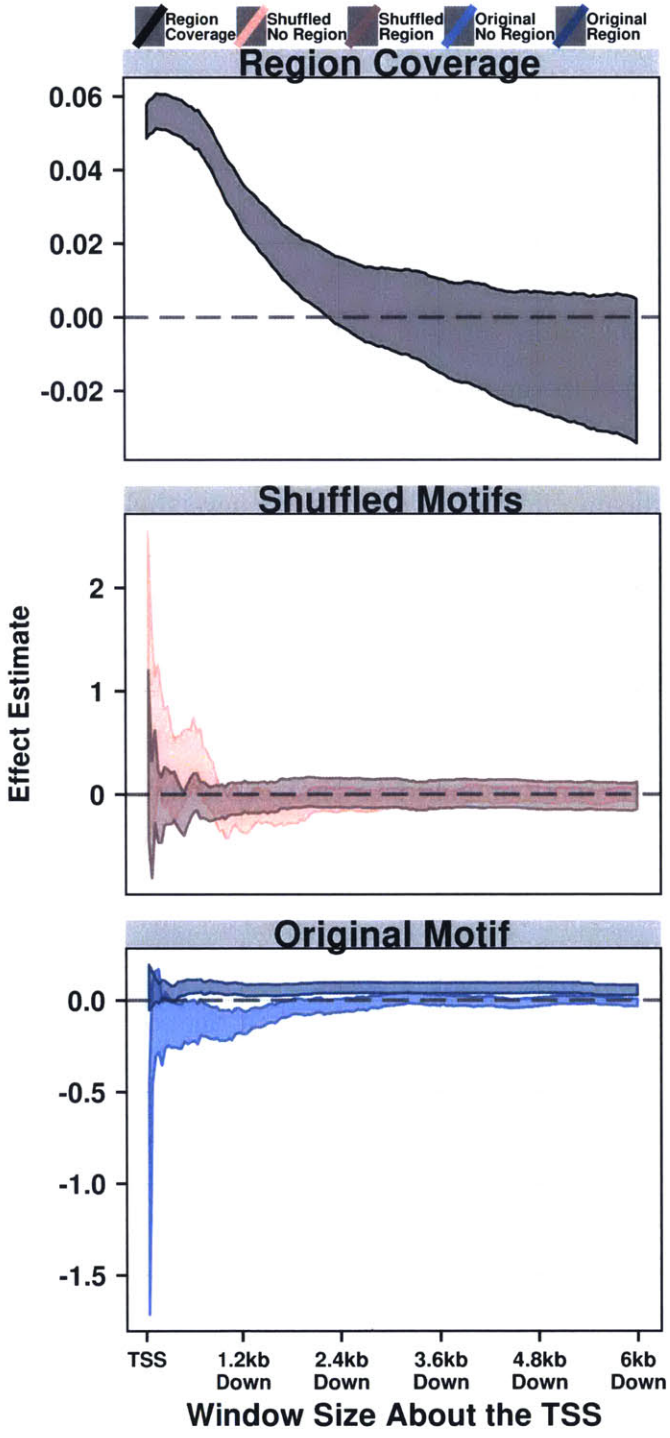

Figure 3-10: Second-order expression effect model coefficient error ranges for the HER-PUD1 motif for various sizes of symmetric TSS windows.

When we repeat this analysis using sliding windows of size 1000 base pairs, we again see strong signs that HERPUD1 motif hits occurring 800 base pairs from TSSs, especially those occurring upstream, have an outsized influence on second-order expression effects (Figure 3-11). Furthermore, the error estimate on this effect indicates that it is statistically significant, while the effect of shuffled motifs is not. Furthermore, only the effect of the original motif overlapping with promoters has a positive effect on coexpression, while original motifs not overlapping with promoters have an almost negligible effect in the other direction, suggesting that if anything they bloack promoter-overlapping sites from being fully effective in mediating regulation.

The most significant effect on coexpression, nevertheless, can be attributed to the presence of promoters in an area extending roughly 1500 base pairs to either side of transcription start sites. This confirms the biological relevance of the promoter regions we are using, as they seem to an absolutely surefire pathway through which HERPUD1 exerts an effect on the expression of its targets. However, the magnitude of the effect of promoters compared to that of HERPUD1 motifs overlapping with promoters is rather small, indicating that while relatively rare and prone to noise, the regulatory effect of these motif hits is still considerable.

We perform a similar analysis with other factors with first-order expression effects to see if the motif hit clusters we identified in Section 2.5 have a direct effect on expression levels. However, as shown in Figure 3-12, this does not seem to be the case. The windows within which these clusters occurred for BHLHE40-known3 and CTCF-known1 do not assign particularly high model values to the original motifs, within and without promoters, relative to their shuffled counterparts and the presence of regulatory regions. Indeed, the coefficient valuen profiles by window location look quite similar in both cases, with shuffled regions having a high estimated effect, suggesting that this model is prone to measuring nucleotide bias with respect to first-order expression effects.

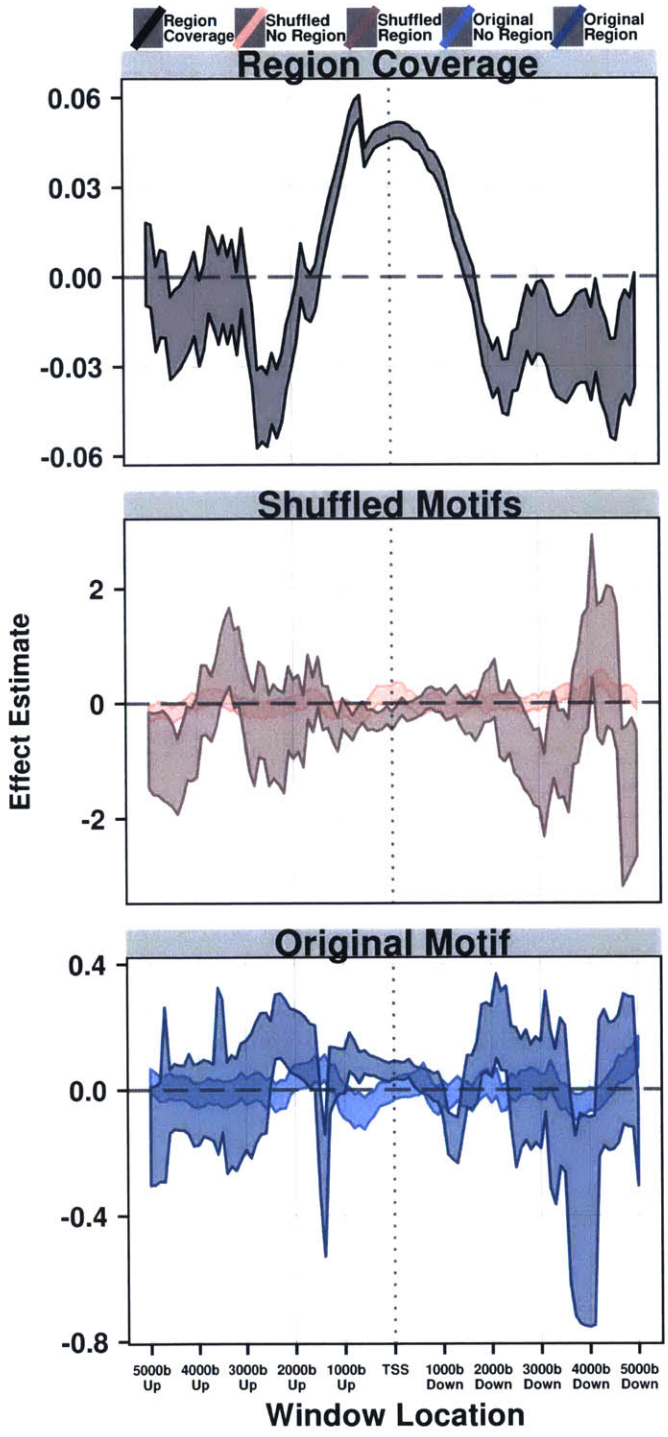

Figure 3-11: Second-order expression effect model coefficient error ranges for the HERPUD1 motif for TSS windows of width 1000bp centred at a range of locations relative to TSSs.

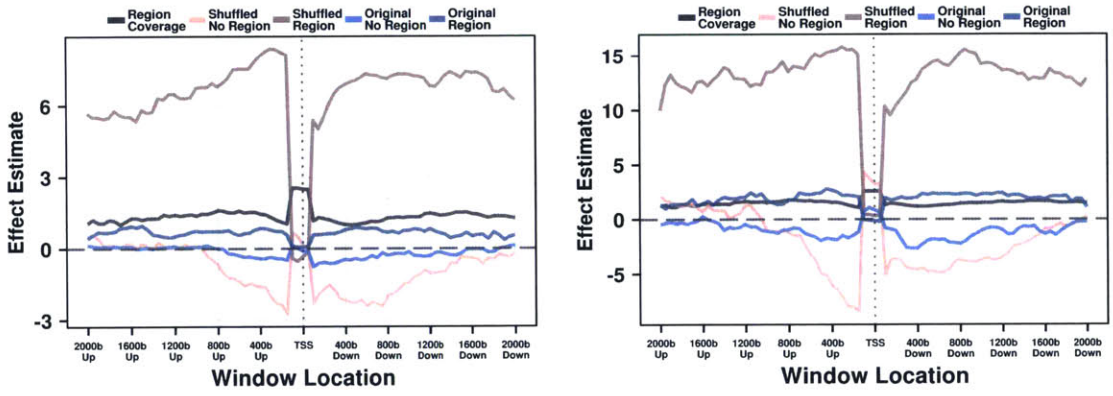

Figure 3-12: Second-order expression effect model coefficient values for BHLHE40-known3 and CTCF-known1 using sliding TSS windows of width 200bp.

Appendix A

Supplementary Information

This appendix contains additional information on how the dataset used in this document were acquired, along with notes on their properties where appropriate. All analysis copleted for this thesis was done using the R statistical language (v3.0.1) running in an `x86-64` linux-GNU environment, with plotting done using the `ggplot2` (v.0.9.3.1), `plyr` (v1.8), and `RColorBrewer` (v1.0-5) packages.

A.1 Expression Dataset Acquisition and Processing

All samples for which RNA-seq transcript data was available were collected the lung tissue subset as defined by the SMTSD field from the GTEx dataset downloaded through dbGap, release date June 13, 2014. This subset was chosen as it was the largest homogenous tissue subset available through GTEx. Transcript expression values were aggregated into RNA product expression values by using the most highly expressed transcript for each of the 57393 products. PCA revealed no significant clustering of this dataset, indicating that the samples were fairly homogenous (Figure A-1).

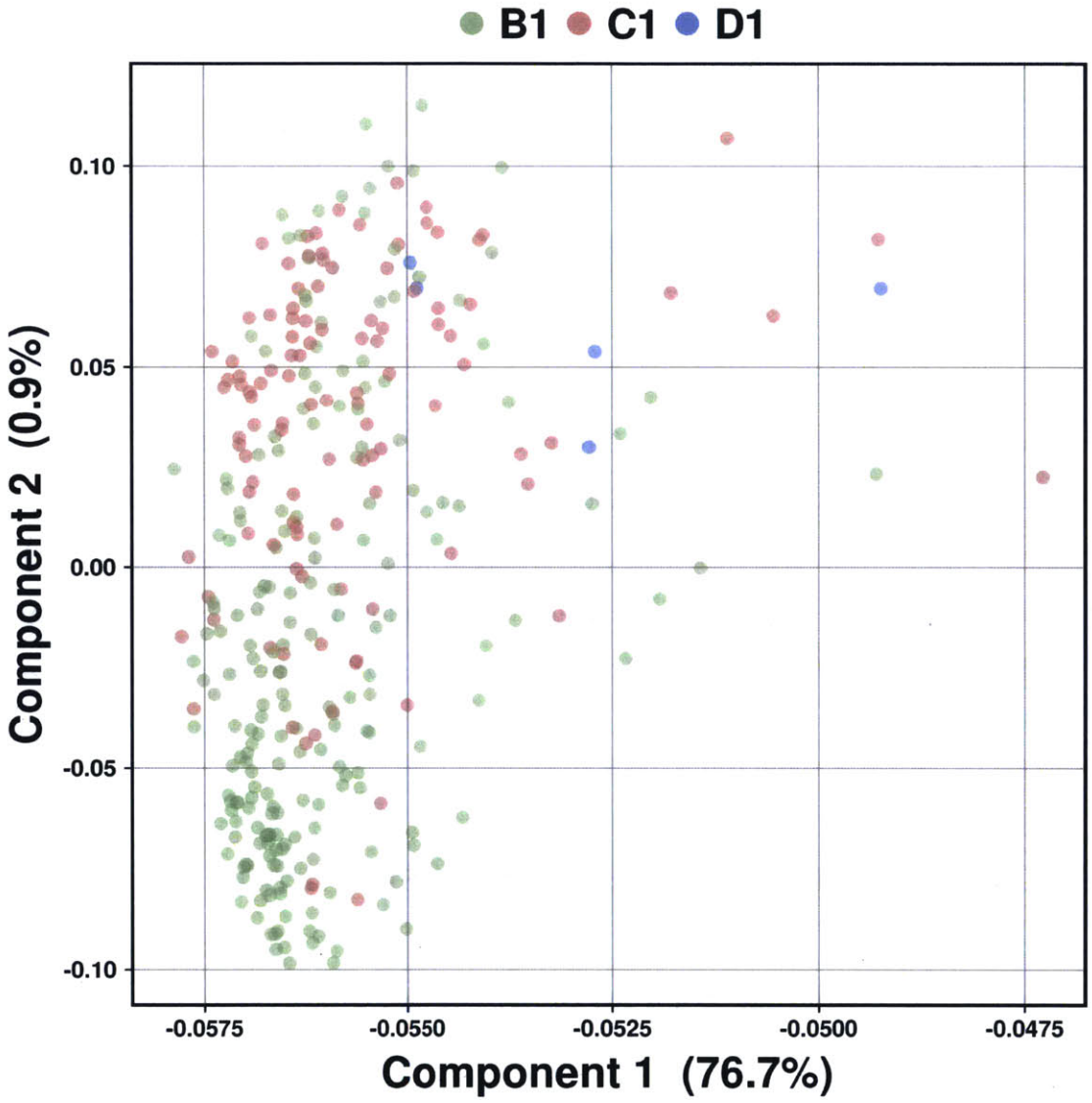

Figure A-1: Principal components analysis of the 320 GTEx lung tissue expression samples, coloured according to the center where the sample was collected.

A.2 Producing Transcription Factor Motif Hits

To identify genomic locations where transcription factors potentially bind, we use the set of TF motifs described by Kheradpour and Kellis (2014) and made publicly available at `http://compbio.mit.edu/encode-motifs/`. For each of the transcription factors analyzed, two types of motifs were included in this set: "known" motifs culled from earlier studies, and "discovered" motifs that were found by scanning peaks in ChIP-seq data made available as part of the ENCODE consortium.

In addition, most motifs have at least one "shuffled" variant, whose use was first introduced in Lindblad-Toh et al. (2011). These are generated by creating 100 random permutations of the positions of the original position weight matrix within blocks defined by information content, and then filtering out permutations that do not have a similar number of genome-wide matches to the original motif. The remaining permutations are then clustered, and one permutation is selected from each cluster to serve as a shuffled motif. The set of shuffled motifs is thus designed to have the same nucleotide make-up as the original motif, and similar higher-order information content patterns, while containing motifs that are diverse in terms of sequence. Some motifs do not have any permutations that pass the match filter, and hence these motifs do not have any shuffled motifs available.

To match motifs to the genome, a custom program developed by Pouya Kheradpour is used. This program finds all matches in a subset of the genome whose hit score is higher than a minimum score corresponding to a given p-value threshold calculated using the TFM-PVALUE algorithm (Touzet and Varré, 2007). The hit score of each match is then normalized using the formula (hit_score - minimum_score)/(max_score - minimum_score). Prior to matching, a pseudocount of 0.001 was added to each position in a motif. Because storing the hits for all motifs genome-wide would take excessive disk space, we limited our scan to regions 10kb upstream of the transcription start sites, regions within the bodies of each of the transcripts used for each gene, and regions 2kb downstream of the end of the final

exon in each of these transcripts.

A.3 Regulatory Region Calls from Reg2Map

Regulatory region calls were downloaded from the Reg2Map website (`http://www.broadinstitute.org/~meuleman/reg2map/`), with the HoneyBadger2 (WM20140520) set of regions at $-\log 10(p) \geq 10$ used. The region calls corresponding to the E096 (Lung) Roadmap epigenetic calls included 17707 regions identified as promoters, 51779 enhancers, and 4392 dyadic regions which can act as either promoter or enhancers.

Bibliography

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K.-k., Dong, X., Djebali, S., Ruan, Y., Davis, C. a., Carninci, P., Lassman, T., Gingeras, T. R., Guigo, R., Birney, E., Weng, Z., Snyder, M., and Gerstein, M. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research*, 22(9):1658–1667.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403.

Ernst, J. and Kellis, M. (2013). Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Research*, 23(7):1142–1154.

Frith, M. C., Li, M. C., and Weng, Z. (2003). Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Research*, 31(13):3666–3668.

Glass, K., Quackenbush, J., Spentzos, D., Haibe-Kains, B., and Yuan, G.-C. (2015). A network model for angiogenesis in ovarian cancer. *BMC Bioinformatics*, 16(1):1–17.

Guo, C., Eckler, M. J., McKenna, W. L., McKinsey, G. L., Rubenstein, J. L. R., and Chen, B. (2013). Fezf2 expression identifies a multipotent progenitor for neocortical projection neurons, astrocytes, and oligodendrocytes. *Neuron*, 80(5):1167–1174.

Janky, R., Verfaillie, A., Imrichová, H., van de Sande, B., Standaert, L., Christiaens, V., Hulselmans, G., Herten, K., Naval Sanchez, M., Potier, D., Svetlichnyy, D., Kalender Atak, Z., Fiers, M., Marine, J. C., and Aerts, S. (2014). iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Computational Biology*, 10(7):e1003731.

Kheradpour, P. and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42(5):2976–87.

Kim, L., Esplugues, E., Zorca, C. E., Parisi, F., Kluger, Y., Kim, T. H., Galjart, N. J., and Flavell, R. a. (2014). Oct-1 Regulates IL-17 Expression by Directing

Interchromosomal Associations in Conjunction with CTCF in T Cells. *Molecular Cell*, 54(1):56–66.

Landt, S. G. and Marinov, G. K. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831.

Lee, T. I. and Young, R. a. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251.

Lercher, M. J., Urrutia, A. O., Pavlíček, A., and Hurst, L. D. (2003). A unification of mosaic structures in the human genome. *Human Molecular Genetics*, 12(19):2411–2415.

Lifanov, A. P., Makeev, V. J., Nazina, A. G., and Papatsenko, D. a. (2003). Homotypic regulatory clusters in Drosophila. *Genome Research*, 13(4):579–588.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Baldwin, J., Bloom, T., Chin, C. W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. a., Cree, A., Dihn, H. H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L. R., Nazareth, L. V., Okwuonu, G., Santibanez, J., Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., Sodergren, E., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard, K. S., Pedersen, J. S., Lander, E. S., and Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–82.

Nachman, I. and Regev, A. (2009). BRNI: Modular analysis of transcriptional regulatory programs. *BMC Bioinformatics*, 10(155).

Nachman, I., Regev, A., and Friedman, N. (2004). Inferring Quantitative Models of Regulatory Networks from Expression Data. *Bioinformatics*, 00(00):1–8.

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. a. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–86.

O'Connor, T. R. and Bailey, T. L. (2014). Creating and validating cis-regulatory maps of tissue-specific gene expression regulation. *Nucleic Acids Research*, 42(17):11000–11010.

Patel, B., Kang, Y., Cui, K., Litt, M., Riberio, M. S. J., Deng, C., Salz, T., Casada, S., Fu, X., Qiu, Y., Zhao, K., and Huang, S. (2014). Aberrant TAL1 activation is mediated by an interchromosomal interaction in human T-cell acute lymphoblastic leukemia. *Leukemia*, 28(2):349–61.

Pique-Regi, R., Degner, J. F., Pai, A. a., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455.

Rao, Y. S., Chai, X. W., Wang, Z. F., Nie, Q. H., and Zhang, X. Q. (2013). Impact of GC content on gene expression pattern in chicken. *Genetics, selection, evolution : GSE*, 45:9.

Sémon, M., Mouchiroud, D., and Duret, L. (2005). Relationship between gene expression and GC-content in mammals: Statistical significance and biological relevance. *Human Molecular Genetics*, 14(3):421–427.

Sherman, M. S. and Cohen, B. a. (2012). Thermodynamic state ensemble models of cis-regulation. *PLoS Computational Biology*, 8(3):1–10.

Shibata, M., Gulden, F. O., and Sestan, N. (2015). From trans to cis: transcriptional regulatory networks in neocortical development. *Trends in Genetics*, 31(2):77–87.

Spitz, F. and Duboule, D. (2008). Global control regions and regulatory landscapes in vertebrate development and evolution. *Advances in Genetics*, 61(07):175–205.

Touzet, H. and Varré, J.-S. (2007). Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms for Molecular Biology*, 2:15.

Warner, J. B., Philippakis, A. a., Jaeger, S. a., Sherry, F., Lin, J., and Bulyk, M. L. (2008). Systematic identification of mammalian regulatory motifs' target genes and functions. *Nature Methods*, 5(4):347–353.