STATISTICAL MODELS FOR THE FUNCTION AND EVOLUTION OF CIS-REGULATORY ELEMENTS IN MAMMALS

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Noah Dukler

May 2019

STATISTICAL MODELS FOR THE FUNCTION AND EVOLUTION OF CIS-REGULATORY ELEMENTS IN MAMMALS

Noah Dukler, Ph.D.
Cornell University 2019

Precise gene regulation is essential for a wide variety of transient, developmental, and homeostatic processes. The majority of gene regulation is mediated by cis-regulatory elements, both distal (enhancers), and proximal (promoters & enhancers). Developments in biochemical assays, gene editing techniques, and sequencing technology have enabled genome-wide profiling of regulatory elements over a wide variety of *in vivo* conditions. In this tripartite work, I present separate statistical frameworks for analyzing how these repertoires of regulatory elements work at both physiological, and evolutionary timescales.

The first part describes the use of PRO-seq to characterize rapid changes in the transcriptional landscape of human cells to celastrol, a compound that has potent anti-inflammatory, tumor-inhibitory, and obesity-controlling effects. By exploiting the ability of PRO-seq to detect nascent RNAs, I characterize the transcriptional response at both genes and enhancers, and leverage statistical models to detect transcription factors that orchestrate it. I implicate several transcription factors in early transcriptional changes, including members of the E2F and RFX families. PRO-seq also allows us to detect an increase in transcription start site proximal pausing, suggesting that pause release may be a mechanism for inhibiting gene expression during the celastrol response. This work demonstrates that a thorough analysis of PRO-seq time-course data can provide novel insight into multiple aspects of a complex transcriptional response.

The second part develops a statistical model for determining whether constituent enhancers of a “super-enhancer” exhibit synergy and thus address the question “Is a super-enhancer greater than the sum of its parts?” In this work I reconcile two works with seemingly opposing theses by finding that we cannot confidently reject synergy-free models for super-enhancers. Furthermore, I demonstrate that thoughtful consideration of null models for synergy in gene regulation is critical for furthering our understanding of ensembles of regulatory elements.

In the final section, I develop evolutionary models for cis-regulatory function as quantified by genome-wide biochemical assays. I apply a noise-aware phylogenetic model to analyze the evolution of H3K27Ac and H3K4me3 histone marks as proxies of enhancer and promoter function. I estimate relative turnover rates for a variety of functional element categories and show that gene expression and sequence constraint correlate with turnover rate. I also propose that dosage sensitivity of target genes can explain the discrepancy between sequence and histone mark turnover rates of associated CREs.

This work illustrates the important role statistical models play in understanding gene regulation at all levels and suggests a potential path towards unified models of gene regulation and evolution.

BIOGRAPHICAL SKETCH

Noah Dukler was born in 1991, near Kingston, NY. He grew up in Gardiner, NY and attended New Paltz High School. Upon graduation from high school in 2009, he attended the State University of New York (SUNY) at Geneseo. He graduated from SUNY Geneseo in 2013 *summa cum laude* with a B.S. in Biochemistry and a B.A. in Mathematics as well as a minor in Computational Biology.

He then was accepted to the Tri-Institutional program for Computational Biology at Cornell University in 2013 where he joined the lab of Dr. Adam Siepel. Upon Dr. Siepel accepting a new position at Cold Spring Harbor Laboratory (CSHL), Noah transfered to Weill Cornell Medicine and did his dissertation work at CSHL as a student in residence. While doing his dissertation work Noah worked on developing models for gene regulation and its evolution in mammalian systems. During this time he was co-first author on a correspondence titled “Is a super-enhancer greater than the sum of its parts?” which was selected as one of the top 10 papers of 2017 at the RECOMB/ISCB Regulatory Systems Genomics conference.

This thesis is dedicated to my family for their constant support, and to Elizabeth Hutton, who kept me sane

ACKNOWLEDGEMENTS

First I would like to acknowledge Dr. Charles Danko, Dr. John Lis, and Dr. Ekta Khurana for agreeing to serve as members of my thesis committee. Their suggestions and insight have been invaluable throughout my career as a graduate student. I would also like to acknowledge the members of the Siepel lab, past and present, for all their help and support these past few years. In particular I would like to thank Brad Gulko for his invaluable assistance in all things bash related, and his invaluable contributions to the super-enhancer project. I would also like to thank Yi-Fei Huang for his patient mentorship without which I would not be here today and Charles Danko for his mentorship while he was a member of the Siepel lab. I would also like to thank the Lis and Danko Labs, in particular Greg Booth, Nate Tippens, and Ed Rice for their experimental work on the Celastrol project. Last, but certainly not least, I would like to thank my PI, Adam Siepel for creating a supportive yet critical environment without which I would not have matured as a scientist.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

NASCENT RNA SEQUENCING REVEALS A DYNAMIC GLOBAL TRANSCRIPTIONAL RESPONSE AT GENES AND ENHANCERS TO THE NATURAL MEDICINAL COMPOUND CELASTROL

Note: With the exception of a few minor changes, this chapter contains the same text as the previously published work “Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol” in Genome Research (Volume 27, Number 11).

1.1 Abstract

Most studies of responses to transcriptional stimuli measure changes in cellular mRNA concentrations. By sequencing nascent RNA instead, it is possible to detect changes in transcription in minutes rather than hours, and thereby distinguish primary from secondary responses to regulatory signals. Here, we describe the use of PRO-seq to characterize the immediate transcriptional response in human cells to celastrol, a compound derived from traditional Chinese medicine that has potent anti-inflammatory, tumor-inhibitory, and obesity-controlling effects. Celastrol is known to elicit a cellular stress response resembling the response to heat shock, but the transcriptional basis of this response remains unclear. Our analysis of PRO-seq data for K562 cells reveals dramatic transcriptional effects soon after celastrol treatment at a broad collection of both coding and noncoding transcription units. This transcriptional response occurred in two major waves, one within 10 minutes, and a second 40-60 minutes after treatment. Transcriptional activity was generally repressed by celastrol, but one distinct group of genes, enriched for roles in the heat-shock response, displayed strong activation. Using a regression approach, we identified key transcription factors that appear to drive these transcriptional responses, including members of the E2F and RFX families. We also found sequence-based evidence that particular TFs drive the activation of enhancers. We observed increased polymerase pausing at both genes and enhancers, suggesting that pause release may be widely inhibited during the celastrol response. Our study demonstrates that a careful analysis of PRO-seq time-course data can disentangle key aspects of a complex transcriptional response, and it provides new insights into the activity of a powerful pharmacological agent.

1.2 Introduction

The technique of perturbing cells and then measuring changes in their patterns of gene expression is a reliable and widely used approach for revealing mechanisms of homeostatic regulation. In mammalian cells, a wide variety of stimuli that induce striking changes in transcription are routinely applied, including heat shock, hormones such as estrogen, androgen and cortisol, lipopolysaccharide, and various drugs. Regardless of the stimulus, transcription is commonly assayed by measuring concentrations of mature mRNA molecules, typically using RNA-seq. This approach is now relatively straightforward and inexpensive, and allows for the use of standard analysis tools in detecting many transcriptional responses(Oshlack et al., 2010; Ozsolak and Milos, 2011).

Nevertheless, these mRNA-based approaches are fundamentally limited in temporal resolution owing to the substantial lag between changes in transcriptional activity and detectable changes in the level of mRNAs. This lag results in part from the time required for transcription and post-transcriptional processing, and in part because pre-existing mRNAs buffer changes in mRNA concentration. For a typical mammalian gene, significant changes may require hours to detect, making it difficult to distinguish primary responses to a signal from secondary regulatory events. A possible remedy for this limitation is instead to make use of GRO-seq (Core et al., 2008), PRO-seq (Kwak et al., 2013), NET-seq (Churchman and Weissman, 2011; Mayer et al., 2015; Nojima et al., 2015), or related methods (Dolken et al., 2008; Rabani et al., 2011; Li et al., 2016) for assaying nascent RNAs. These assays have the important advantage of directly measuring the production of new RNAs, rather than concentrations of mature mRNAs. As a consequence, they can detect immediate changes in transcriptional activity, and they permit time courses with resolutions on the order of minutes rather than hours (Danko et al., 2013; Hah et al., 2011; Jonkers et al., 2014; Mahat et al., 2016). An additional benefit of nascent RNA sequencing is that it is effective in detecting unstable noncoding RNAs, including enhancer RNAs (eRNAs), together with protein-coding transcription units (Hah et al., 2011, 2013; Core et al., 2014). As a result, both active regulatory elements (which are generally well marked by eRNAs) and transcriptional responses can be detected using a single assay (Danko et al., 2015).

In this study, we sought to use PRO-seq to characterize the immediate, dynamic transcriptional response to the compound celastrol. Celastrol (also known as tripterine) is a pentacyclic triterpenoid isolated from the root extracts of Tripterygium wilfordii (thunder god vine), which has been used for millennia in traditional Chinese medicine for treatment of fever, joint pain, rheumatoid arthritis, bacterial infection, and other ailments (Westerheide et al., 2004). During the past few decades, celastrol has shown promise as an anti-inflammatory agent in animal models of collagen-induced arthritis, Alzheimer's disease, asthma, systemic lupus erythematosus, and rheumatoid arthritis (Guo et al., 1981; Allison et al., 2001; Xu et al., 2003; Li et al., 2005; Sethi et al., 2007). In addition, celastrol is known to inhibit the proliferation of tumor cells, including those from leukemia, gliomas, prostate, and head/neck cancer (Sethi et al., 2007; Nagase et al., 2003; Yang et al., 2006; Fribley et al., 2015). Recent research has also demonstrated striking obesity-controlling effects in mice (Liu et al., 2015; Ma et al., 2015).

Celastrol is known to activate the mammalian heat shock transcription factor HSF1 and stimulate the heat shock response(Westerheide et al., 2004; Trott et al., 2008) as well as the unfolded protein response (Fribley et al., 2015; Mu et al., 2008). In addition, celastrol activates a battery of antioxidant response genes (Trott et al., 2008). At the same time, celastrol inhibits the activities of other transcription factors, including androgen receptor (Hieronymus et al., 2006), glucocorticoid receptor (Trott et al., 2008) and NF-KB (Sethi et al., 2007). Thus, in several respects, the cellular response to celastrol resembles other well-known stress responses, including, in particular, the response to heat shock. Indeed, this heat-shock-like behavior appears to explain, in part, the cytoprotective properties of celastrol. Nevertheless, it remains unclear exactly what distinguishes the celastrol response from heat shock and other stress responses. In particular, little is known about the immediate transcriptional effects or primary targets of celastrol. Thus, an examination using PRO-seq provides an opportunity for a deeper understanding of the specific mechanisms underlying the activity of this potent compound, with potential therapeutic implications.

With these goals in mind, we collected PRO-seq data for K562 cells at tightly spaced time points after treatment with celastrol and analyzed these data using a variety of computational methods. Our analysis sheds new light on the immediate transcriptional response to celastrol at both regulatory elements and target genes. More generally, it demonstrates that time-courses of PRO-seq data together with appropriate bioinformatic analyses can be used to dissect key aspects of a complex transcriptional response.

1.3 Results

1.3.1 Celastrol induces broad transcriptional repression and more limited up-regulation

We prepared PRO-seq libraries for K562 cells before celastrol treatment and after 10, 20, 40, 60, and 160 minutes of celastrol treatment, with two biological replicates per time point (Figure 1.1A). To ensure that we could normalize read counts even in the presence of global changes in transcription, we spiked the same number of permeable Drosophila cells into each sample prior to run-on (Booth et al. 2016). Samples were sequenced to a total combined depth of 334.3M reads, with an average replicate concordance of r^2 = 98% (Supplemental Figure 1.8). About 0.5M of these reads (0.1%) were derived from the Drosophila spike in. To obtain gene models appropriate for our cell types and conditions, we developed a probabilistic method, called tuSelector, that considers all GENCODE-annotated isoforms for each gene and identifies the most likely gene model given our PRO-seq data (Supplemental Figure 1.9, Supplementary Methods). This step was particularly important for analyses that de-

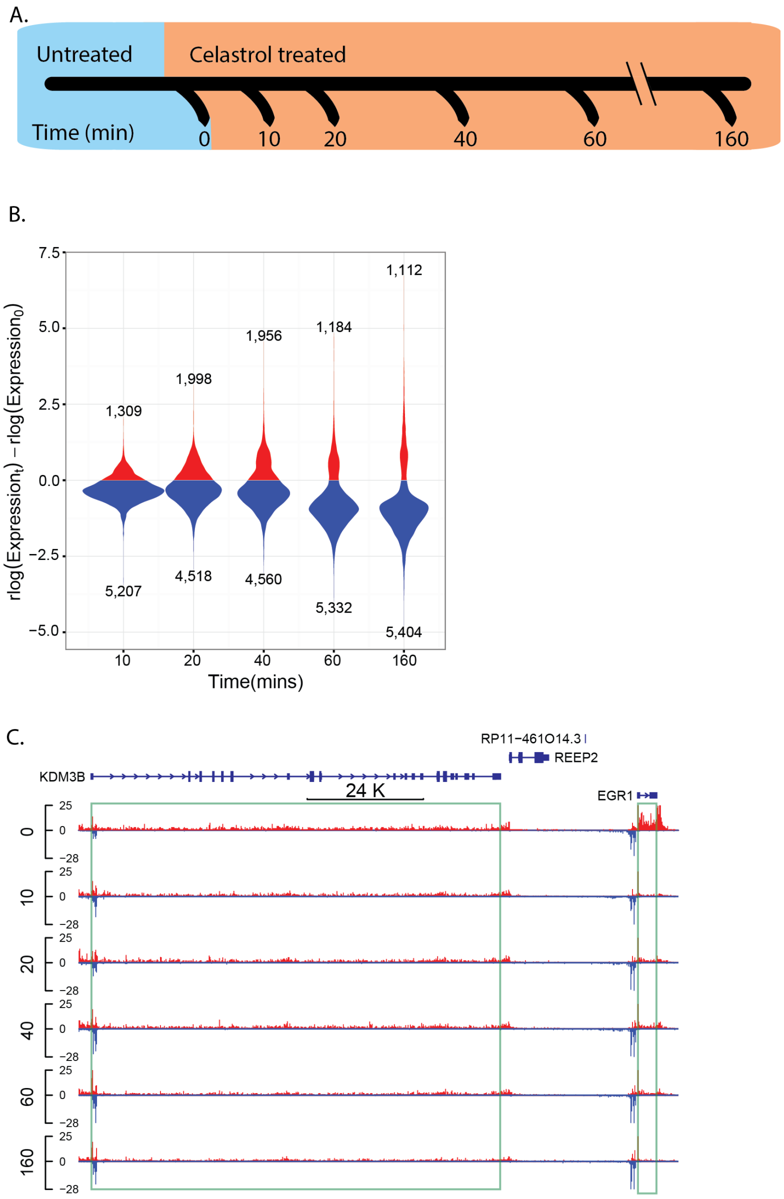


Figure 1.1: **Characterizing the dynamic transcriptional response to celastrol using PRO-seq.** (A) PRO-seq was applied to K562 cells collected before celastrol treatment (untreated / 0 minutes) and at 10, 20, 40, 60, and 160 minutes after celastrol treatment. Two biological replicates were performed for each time point. (B) Distribution of log expression ratios (treated vs. untreated) for each time point (rlog is a regularized log2 estimate obtained from DESeq2). Only genes classified as differentially expressed (DE) throughout the time course are represented. Notice that most DE genes (FDR ≤ 0.01) are down-regulated upon celastrol treatment. (C) A UCSC Genome Browser display showing raw PRO-seq data for two differentially expressed genes, EGR1 and KDM3B. EGR1 is rapidly and strongly repressed (immediate decrease of ∽ 80%), whereas KDM3B is more gradually repressed, losing ∽ 50% of its expression by 160 minutes.

pend on an accurate transcription start site for each gene. We identified a total of 12,242 protein-coding genes from GENCODE as being actively transcribed in one or more time-points (Methods). Of these genes, 75.4% were active across all six time points, 11.7% were active in a single time-point, and the remaining 12.9% were active in 2-5 time-points. Thus, our PRO-seq data and computational analyses indicate that more than half of all protein-coding genes are transcribed either in the basal condition or during the celastrol response in K562 cells.

The genes that are differentially transcribed in response to celastrol were of particular interest for further analysis. To measure transcriptional activity specific to each time point, we used counts of PRO-seq reads mapping to the first 16 kb of each gene body, omitting the first 500 bp to avoid the effects of promoter-proximal pausing. Because RNA polymerase travels at an average rate of 2 kb/min (Singh and Padgett 2009; Danko et al. 2013; Jonkers et al. 2014; Veloso et al. 2014) and our time points are separated by at least 10 minutes, this strategy conservatively considers new transcription only, yet maintains sufficient statistical power for downstream analysis (see Methods). By applying DESeq2 (Love et al. 2014) to these 16-kb read counts, we identified 6,516 (56%) of the active genes as being differentially expressed (DE) relative to the untreated condition (FDR $\leq$ 0.01). Interestingly, $\backsim$ 80% of these DE genes were down-regulated. Many genes showed rapid and dramatic down-regulation, with decreases in expression by half or more at 3.5% of DE genes within 10 minutes, at 7.8% of DE genes within 20 minutes, and at 48.1% of DE genes within 160 minutes (1.1 1B&C). By contrast, many fewer genes showed substantial increases in expression; for example, only 0.03%, 1.9%, and 7.7% of DE genes had doubled in expression after 10, 20, and 160 minutes, respectively. Nevertheless, extreme up-

and down-regulation were both rare, with < 1% of DE genes showing increases and < 1% showing decreases in transcription by factors of eight or more. These observations are reminiscent of findings for the heat shock response, which have included general decreases in transcription together with up-regulation of selected stress-response elements (Hieda et al. 2004; Mahat et al. 2016b), but the effect of celastrol is somewhat less dramatic. We conclude that celastrol broadly inhibits transcription within minutes after administration but also rapidly activates a set of genes that may be important for continued cellular viability.

1.3.2 Celastrol activates heat shock more strongly and directly than it activates the unfolded protein response

Celastrol has been reported to activate stress response pathways such as the heat shock and unfolded protein responses (Westerheide et al. 2004; Trott et al. 2008; Mu et al. 2008; Fribley et al. 2015). To see whether these effects were detectable at the transcriptional level immediately after treatment with celastrol, we examined our PRO-seq data at genes activated by heat shock factor protein 1 (HSF1) and genes involved in the three branches of the unfolded protein response (UPR), corresponding to activating transcription factor 6 (ATF6), inositol-requiring enzyme 1 (ERN1), and endoplasmic reticulum kinase (EIF2AK3) (Figure 1.2A). Because the initial stages of the UPR and heat shock response are non-transcriptional, we looked for downstream activity of the first group of transcription factors activated in each pathway, using targets reported in the Reactome pathway database (Fabregat et al. 2016). Most direct targets of HSF1 were up-regulated within 160 minutes (Figure 1.2B). Genes en-

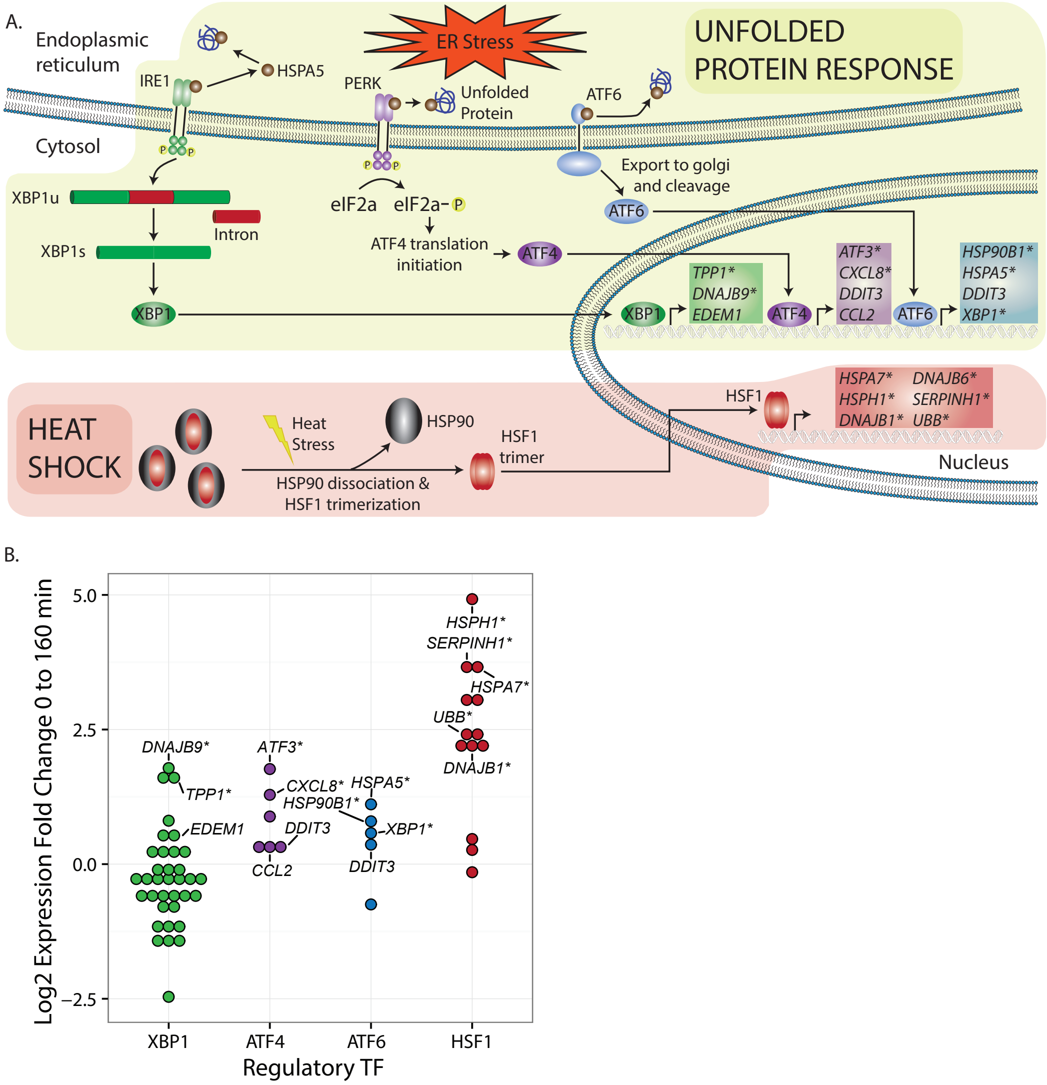


Figure 1.2: **Induction of cellular stress responses by celastrol.** (A) Illustration showing key aspects of the unfolded protein response (UPR) and heat shock response (HSR), both of which have been reported to be induced by celastrol (Mu et al. 2008; Trott et al. 2008). Expected transcriptional targets are shown inside the nucleus, with targets of HSF1, the key transcription factor (TF) activated in the HSR, in red, and targets of the TFs associated with the three major branches of the UPR-XBP1, ATF4, and ATF6-in green, purple, and blue, respectively. Asterisks indicate genes that were differentially expressed in our experiments with FDR ≤ 0.01. (B) PRO-seq-based log fold changes in expression in K562 cells after 160 minutes of treatment by celastrol for numerous known targets of the same four TFs: XBP1, ATF4, ATF6, and HSF1. Genes labeled are the same as those listed in (A). Only targets of HSF1 display strong up-regulation.

coding chaperone protein HSPH1 and proteinase inhibitor CBP1 were among the HSF1 targets showing the strongest initial response, with the gene encoding HSPH1 almost quadrupling its expression in 10 minutes and that for SERPINH1 increasing more than eight-fold in 160 minutes. Most direct targets of the main UPR TFs (ATF4, ATF6, and XBP1), however, were not strongly induced during our time course. There were some exceptions to this general rule, such as genes encoding transcription factor ATF3, chaperone HSPA5, and apoptosis inhibitor DNAJB9, which more than doubled in expression. It is possible that these targets are activated earlier than other targets, perhaps by other TFs. In any case, our observations suggest that celastrol induces a pronounced, rapid transcriptional response in the heat shock pathway, and has a much less pronounced transcriptional effect on the UPR, although some targets of the UPR are activated.

Celastrol produces distinct temporal patterns of transcriptional response

Our PRO-seq data for closely spaced time points enabled us to examine the temporal patterns of transcriptional response to celastrol treatment across the genome. To group our ~6,500 DE genes by shared transcriptional trajectories across the five time points following celastrol treatment, we used the autoregressive clustering algorithm, EMMIX-WIRE, which considers both expression values at each time point and the correlation of these values over time (Wang et al., 2012). EMMIX-WIRE identified four clusters of DE genes showing distinct patterns of transcription (see Methods; Figure 1.3A). Only one of these clusters (cluster #1) displayed dramatic and sustained up-regulation. In contrast, cluster #2 showed rapid and pronounced down-regulation, cluster #3

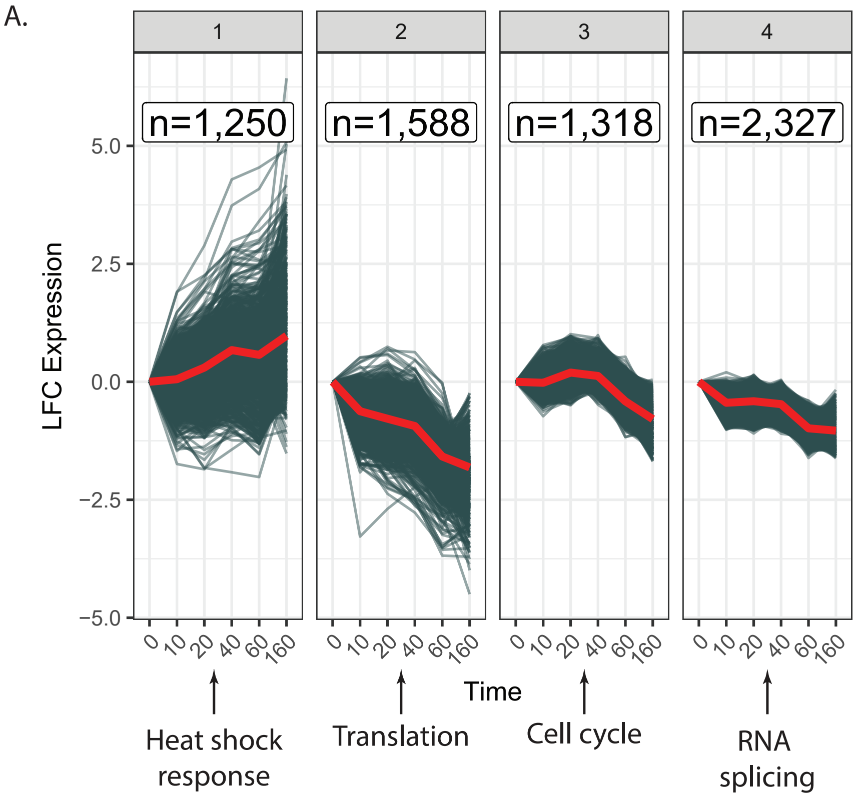


Figure 1.3: **Clusters of genes showing distinct temporal patterns of response to celastrol.** (A) Differentially expressed genes (FDR ≤ 0.01) clustered by time series of log2 fold change (LFC) in expression relative to the untreated condition (0 minutes). Each gene is represented by a blue line, and the red lines indicate the mean expression per time-point per cluster. Below each cluster is a summary of the enriched terms in the Reactome ontology (Fabregat et al. 2016) (FDR ≤ 0.01; see Supplementary Material for details). (B) ChIP-seq data from Vihervaara et al. (2013) describing binding of HSF1 in K562 cells under normal (left) and heat-shock (right) conditions, stratified by our cluster assignments. Each line represents an average over all genes in the cluster in the region of the TSS, with lighter-colored bands representing 95% confidence intervals obtained by bootstrap sampling. Notice that cluster one is unique in showing a strong enrichment for heat-shock-induced binding of HSF1.

showed delayed down-regulation, and cluster #4 displayed moderate, continuous down-regulation. Interestingly, the expression patterns for these clusters suggested that the transcriptional response to celastrol occurs largely in two distinct waves: one within the first ten minutes, and a second between 40 and 60 minutes after treatment. It is possible, however, that additional waves occur but are undetectable at the resolution of our time points. These findings were robust to the number of clusters selected, with similar overall behavior for five-, six-, and seven-cluster models (Supplemental Figures 1.11-1.13).

Each of these clusters was enriched for genes with a distinct biological function, according to Reactome, a carefully curated database of gene-pathway relationships (Figure 1.3A). To identify these enrichments, we labeled genes with their associated pathways from Reactome, then used permutation testing to find pathways overrepresented in each of the four clusters relative to the other three. Cluster #1 is enriched for genes responsible for the HSF1 response, including the HSPA family (Supplemental Figure 7). Consistent with this observation, genes in this cluster have been shown, by ChIP-seq, to bind by HSF1 under heat shock conditions in K562 cells (Figure 1.3B) (Vihervaara et al., 2013). Cluster #2 is enriched for a wide variety of terms corresponding to ribosomal assembly, translational initiation, and peptide elongation. The pronounced transcriptional repression of this cluster, together with activity of HSPB1 (HSP27) and HSPA2 (HSP70), is consistent with reports that celastrol activates the heat shock response and thereby inhibits translation via HSPB1 in the absence of HSP70 (Cuesta et al. 2000) (Supplemental Figure 1.15,1.16). Cluster #3 is enriched for pathways that enable DNA replication (e.g., MCM family) and cell cycle progression (e.g., CDK family; Supplemental Figure 1.17). The delayed down-regulation of these genes may occur as the cell is preparing to enter replicative arrest and, potentially, senescence, consistent with observations that celastrol induces cell cycle arrest and potentiates apoptosis (Feng et al., 2013; Fribley et al., 2015; Kannaiyan et al., 2011). This transcriptional pattern is also consistent with our observation that celastrol-treated cells failed to replicate and mostly died within three days (Supplemental Figure 1.18). Finally, cluster #4 contains essential elements of the RNA splicing machinery (e.g., CD2BP2, CLP1, and the SRSF kinase family) (Supplemental Figure 1.18). Down-regulation of this cluster is consistent with previous reports that splicing is inhibited under heat shock (Shalgi et al., 2013). Similar patterns of enrichment were observed with five-, six-, and seven-cluster models. Overall, these results demonstrate that the components of a complex, multi-layered transcriptional response can be disentangled, to a degree, by identifying groups of genes that display distinct temporal patterns of gene expression.

Several key transcription factors are associated with the celastrol response

Can the distinct transcriptional responses in these clusters of genes be traced to particular transcription factors (TFs)? To address this question, we used linear regression to explain the estimated transcription levels at each time point based on the TFs that apparently bind in the promoter region of each gene (Figure 1.4A). We used two orthogonal sources of information about TF binding: (1) ChIP-seq peaks for untreated K562 cells (Dunham et al., 2012); and (2) scores from DeepBind, a machine-learning method that is trained on a combination of ChIP-seq and in-vitro data and predicts TF-specific binding affinity based on the DNA sequence alone (Alipanahi et al., 2015). In both cases, we considered the interval between 500bp upstream and 200bp downstream of the transcription

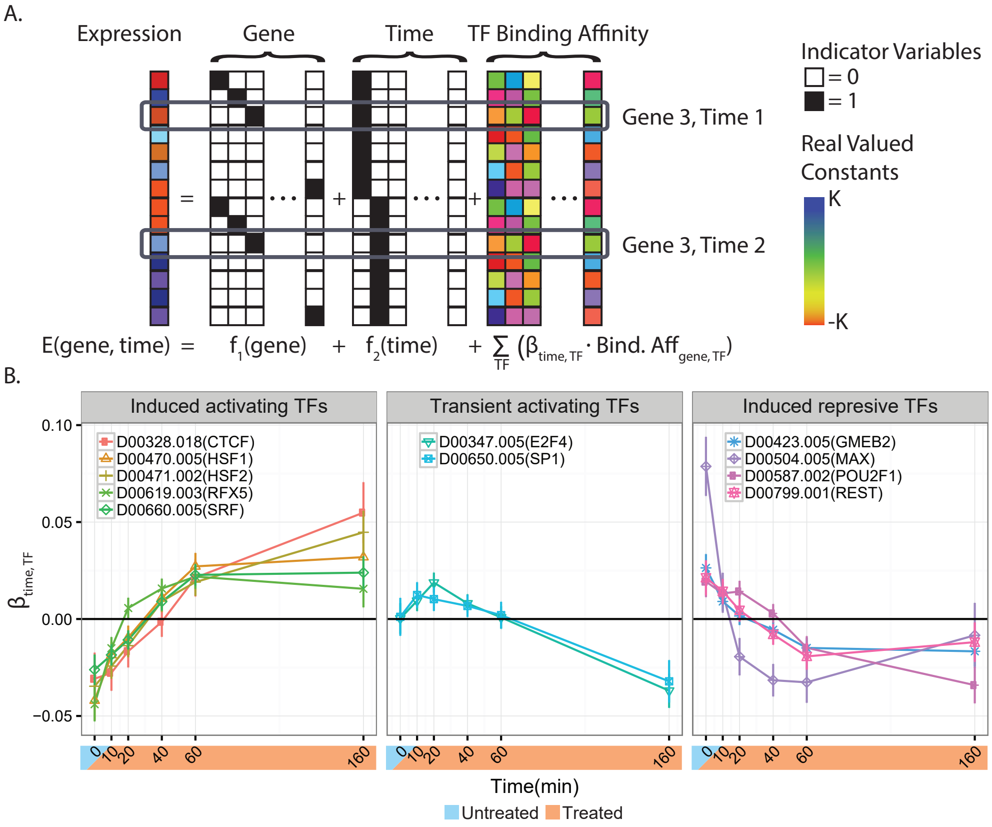


Figure 1.4: **The celastrol response appears to be influenced by various transcription factors (TFs).**(A) A schematic representation of the regression model used to estimate the impact of TF binding. The predicted expression level for each gene is a linear function of a gene-specific term, a time-point specific term, and a sum of per-time TF-specific effects weighted by the estimated relative TF binding affinity in the promoter regions of differentially expressed genes (from DeepBind or ChIP-seq). Note that the TF binding affinity is estimated from the untreated state and is invariant across time. (B) The estimated TF- and time-specific coefficients from the DeepBind-based regression mode. Error bars correspond to the 95% CI estimates for each coefficient. Each gene (e.g., HSF1) is labeled by its DeepBind motif (e.g., D00470.005) in the legend. Only TFs with at least one significant time point (FDR ≤ 0.01) and an absolute effect size in the 90th percentile or above are shown. A positive weight for a TF at a given time-point indicates that genes at which that TF is predicted to be bound showed increased expression relative to those without it. Negative weights indicate decreased expression. The three separate plots represent manually selected clusters of TFs associated with distinct temporal patterns. The time-point-specific TF coefficients explain ∽ 11% of the residual variance not explained by gene-specific or time-point-specific terms.

start site of each active gene. Our regression model included a coefficient for each TF at each time point. A positive estimate of this coefficient indicated that increased affinity for a TF was associated with increased expression at a given time point, whereas a negative estimate indicated that increased affinity for a TF was associated with decreased expression at that time point.

Between the two TF binding datasets, we identified over twenty TFs as being significantly associated with changes in gene expression and having a large effect size (Figure 1.4B, Supplemental Figure 1.20). Of these TFs, E2F4 stood out as showing a particularly pronounced impact on expression in both datasets. E2F4 is associated with incrementally increased expression between 0 and 60 minutes, and with decreased expression thereafter, similar to the expression pattern of genes in cluster #3, which are associated with cell cycle control. This observation is consistent with reports that E2F4 is an activator in some contexts but primarily acts as a repressor responsible for maintaining G2 arrest (Lee et al., 2011; Polager and Ginsberg, 2003). In addition, we found that the dimerizing TFs MYC (from ChIP-seq data) and MAX (from DeepBind predictions), were both associated with an immediate increase in gene expression, followed by decreased expression within 20 minutes (Figure 1.4B). This delayed decrease in expression of MYC- and MAX-bound genes could result from the known disruption of MYC-MAX dimerization by celastrol (Westerheide et al., 2004; Wang et al., 2015a). Finally, genes predicted to be bound by SRF also displayed elevated gene expression after 40 minutes. SRF has been previously associated with early and transient induction of cytoskeletal genes in response to heat stress in murine embryonic fibroblasts (Mahat et al., 2016).

We also found that the paralogous TFs RFX1 (from ChIP-seq) and RFX5 (from DeepBind) were both associated with increases in expression. Both of these TFs have been implicated in regulating the expression of immunity-related human leukocyte antigen (HLA) genes and both have context-specific transcriptional repression and activation mechanisms (Katan et al., 1997; Villard et al., 2000; Xu et al., 2006), so it is possible that they contribute to celastrol's anti-inflammatory effects. However, RFX1 was not tested with DeepBind and its motif is quite similar to that of RFX5, so it is impossible to know from our data whether one or both of these TFs are important in the celastrol response (notably, they do have different binding patterns in vivo in the untreated condition; Supplemental Figure 1.20B,C). Nevertheless, our regression framework is useful in providing a list of candidate TFs whose binding preferences correlate with aspects of the celastrol response.

1.3.3 Increased polymerase pausing is broadly associated with transcriptional repression

Promoter-proximal pausing of RNA polymerase is a rate-limiting and independently regulated step in productive transcription (Andrulis et al., 2000; Wu et al., 2003). Notably, the peaks of paused RNA polymerase at DE genes doubled in height during our time course (Figure 1.5A, Supplemental Figure 1.21). Accordingly, we found that the "log pause index," or log2 ratio of average read depth at the pause peak to that in the proximal gene body, increased by more than 1 (corresponding to a fold-change of more than 2 in the pause index) in DE genes by 160 minutes (Figure 1.5B). Together, these observations indicate that most DE genes undergo increased pausing after celastrol treatment, suggesting that

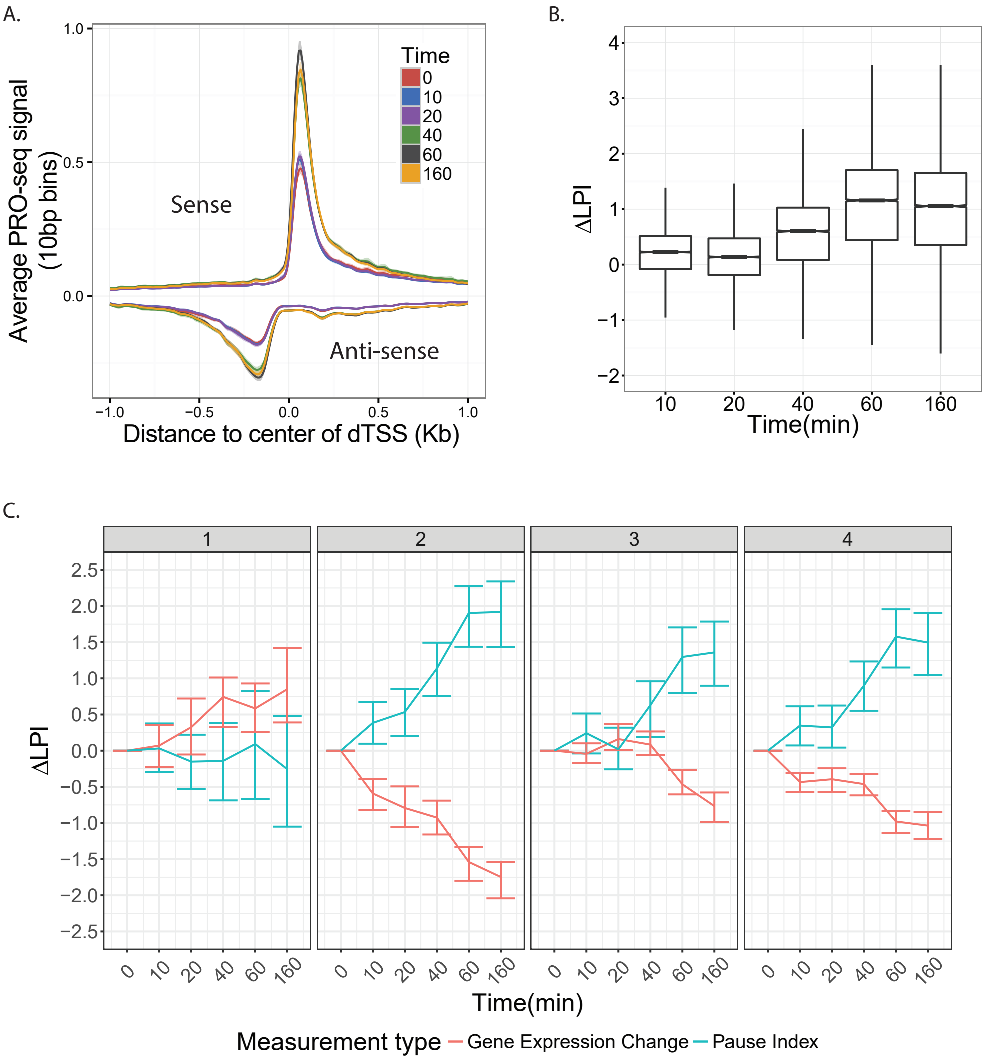


Figure 1.5: **Increased promoter-proximal pausing is associated with transcriptional repression in the response to celastrol.** (A) Mean PRO-seq signal at promoters for all active genes, grouped by time point and oriented with respect to the direction of transcription of the gene. X-axis represents distance to the center of the divergent transcription start site (see Methods). Intervals around each line represent 95% confidence intervals obtained by bootstrap sampling. Notice the general increase in the height of the pause peaks with time. (B) The distribution of changes in the log fold index with respect to the untreated condition (ΔLPI; see Methods) for all active genes at each time point. The notch corresponds to median $+/- 1.58 \cdot IQR/\sqrt{n}$, roughly a 95% confidence interval of the median. (C) The distribution of ΔLPI for all DE genes (FDR $\leq$ 0.01) by cluster and time-point. Notice that all clusters show an increase in the pause index with time, except for cluster #1. Error bars indicate the 25th and 75th percentiles of the data.

pause release may be widely inhibited during the celastrol response.

To see if particular expression patterns were associated with changes in pausing, we separately examined the changes in log pause index during the time course for each of our six gene expression clusters. Interestingly, we found that pausing increased in all clusters with the exception of cluster #1 (Figure 1.5C), the only strongly up-regulated cluster (see Figure 1.3A), where pausing decreased for the majority of genes, but increased or remained unchanged for a significant fraction of them (Supplemental Figure 1.21B). Thus, changes in the log pause index are generally negatively correlated with changes in expression across clusters. This observation suggests that decreases in the rate of release of paused Pol II to productive elongation could contribute to increased pausing and, hence, to down-regulation of transcription, while the absence of such an effect (in cluster #1) might permit up-regulation of transcription (Mahat et al., 2016; Zeitlinger et al., 2007). As cluster #1 is strongly associated with the HSF1 response, this finding is consistent with previous reports that HSF1 regulates transcription by increasing the rate of release of paused RNA polymerase into productive elongation (Mahat et al., 2016). Nevertheless, it is also possible that the inverse correlation between pausing and expression is a consequence of "mass action" of available Pol II (see Discussion).

1.3.4 Heat shock induces a similar, but more pronounced, transcriptional response than celastrol

Celastrol is known to mimic heat-shock in many respects (Westerheide et al., 2004), but it remains unclear how similar the transcriptional responses to these

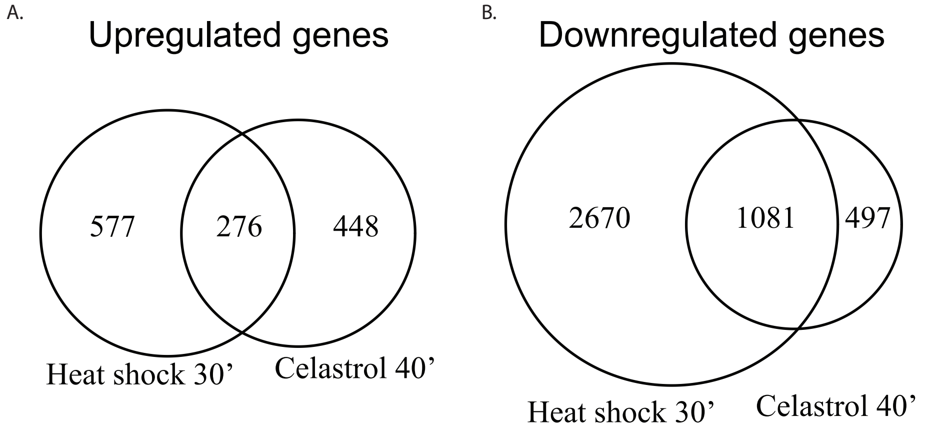


Figure 1.6: **Celastrol downregulates most of the same genes as heat shock, but upregulates many different genes.** (A) Venn diagram of genes that are up-regulated after 30 minutes of heat shock vs. after 40 minutes of celastrol treatment. (B) Venn diagram of genes that are down-regulated after 30 minutes of heat shock vs. after 40 minutes of celastrol treatment.

two stimuli are. To address this question more directly, we obtained PRO-seq data for heat-shock treated K562 cells from a recently published study (Vihervaara et al., 2017) and processed it identically to our celastrol data. We focused on comparing the heat-shock data for 0 and 30 minutes (the only time points available) with our celastrol data for 0 and 40 minutes, additionally considering our 60- and 160-minute time-points for some analyses (Supplemental Figure 1.22A&B). In general, heat shock induced a broader response than celastrol treatment, with twice as many genes differentially expressed after 30 minutes of heat shock (4,604) vs. after 40 minutes of celastrol treatment (2,302). Of the 1,301 genes that were up-regulated in response to either treatment, 21% were shared between the heat shock and celastrol responses, and of the 4,230 that were down-regulated in response to either treatment, 25% were shared (Figure 1.6). As with celastrol treatment, the pause index significantly increased after 30 minutes of heat shock. Taken together, these results suggest that there are many commonalities between the early transcriptional responses to the heat shock and celastrol treatments, but also many differences.

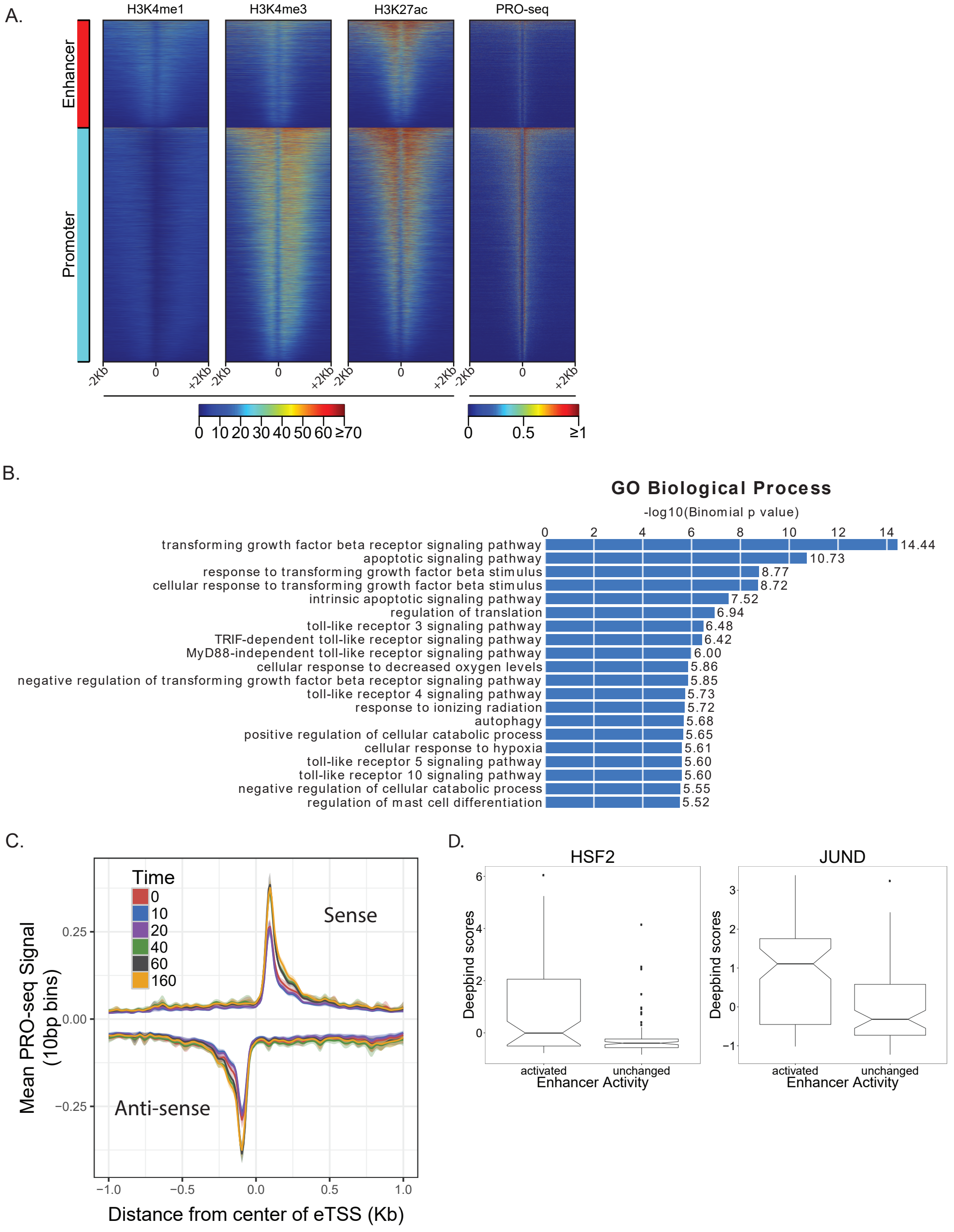
We then sought to characterize the pathways underlying major differences between the celastrol and heat shock responses. Using Reactome, we tested for functional enrichments among shared and non-shared DE genes, separately considering up- and down-regulated genes. We found that genes down-regulated only in the celastrol response were enriched for mitochondrial energy production and translation of mitochondrial genes (Supplemental Figure 1.23A). By contrast, down-regulated genes specific to the heat-shock response were enriched for MAP kinase signaling and cell-cycle progression (Supplemental Figure 1.23B), whereas down-regulated genes that were shared in both responses were strongly enriched for ribosomal formation and translation (Supplemental Figure 1.23C). For the up-regulated genes, the heat-shock-only genes were enriched for GPCR signaling, and the shared genes were dominated by heat-shock response pathways in agreement with the analyses discussed above (Supplemental Figure 1.24A&B); no pathways were significantly enriched in the celastrol-only response.

Finally, we searched for pathways whose genes tended to change expression in opposite directions in the heat shock and celastrol responses. We compared the 40, 60, and 160-minute time-points for celastrol to the 30-minute heat-shock time-point. One pathway, cholesterol biosynthesis, emerged from this analysis as down-regulated in celastrol at both 60 and 160 minutes but up-regulated in heat-shock (Supplemental Figure 1.25A). This observation is consistent with previous findings for mammalian cells that heat shock increases activity of the MVA pathway, a key cholesterol biosynthesis pathway (Shack et al., 1999). The central regulatory enzyme in the MVA pathway, HMGCR, is clearly up-regulated in heat shock and down-regulated at 60 minutes in celastrol. We investigated whether sterol response element binding factor 1 (SREBF1), an important TF for cholesterol biosynthesis genes (Brown and Goldstein, 1997), was a potential mechanism for decreased genic expression in the celastrol response by asking if genes bound by SREBF1 in untreated cells showed decreased expression relative to those not bound across the celastrol time-course (Supplemental Figure 1.25B). We found that genes that were strongly bound by SREBF1 in untreated cells, in comparison to unbound genes, went from being more highly expressed in the untreated condition, to having similar mean expression at 60 minutes of treatment, and lower mean expression at 160 minutes of treatment. These results demonstrate that despite having many similar effects, celastrol and heat shock have opposite effects on the expression of genes involved in cholesterol biosynthesis.

1.3.5 Enhancers show similar functional associations and pausing patterns to genes

Previous studies have shown that putative enhancers are divergently transcribed, producing nascent RNAs that can be detected via PRO-seq (Danko et al., 2015; Core et al., 2014; De Santa et al., 2010). Using dREG (Danko et al., 2015), which predicts divergent transcription start sites (dTSS) from stranded GRO/PRO-seq data, we identified 25,891 apparent dTSS from our PRO-seq data, pooling calls across time points. Based on the distance from nearest annotated genic TSS, we classified 7,334 of these dTSS as likely transcribed enhancers, 15,941 as likely promoters, and the remaining 2,616 as ambiguous. For

Figure 1.7: **Response to celastrol at predicted transcribed enhancers.** (A) Divergent transcription start sites (dTSS) that were classified by distance-based rules as likely enhancers or promoters show distinct patterns of histone marks in untreated K562 cells. Shown are H4K4me1 (enriched at enhancers), H3K4me3 (enriched at promoters), and H3K27ac (enriched at both). PRO-seq read counts are shown for comparison. X-axis is oriented by direction of transcription of nearest gene. (B) Gene Ontology (GO) biological processes associated with differentially expressed enhancers using GREAT (McLean et al. 2010). Bar plot represents –log10 p-values for enrichment, with numerical fold enrichments indicated at right. (C) Metaplot of PRO-seq signal at all enhancers, centered on the dTSS, per time-point. Units of PRO-seq signal are average numbers of reads per 10bp bin. Intervals around each line represent a 95% confidence interval obtained by bootstrap sampling. (D) The distribution of DeepBind scores for HSF2 and JUND for transcriptionally activated and unchanged sets of enhancers. The notch corresponds to the median $+/- 1.58 \cdot IQR/\sqrt{n}$, roughly a 95% confidence interval of the median.


A.
H3K4me1
H3K4me3
H3K27ac
PRO-seq
Enhancer
Promoter
-2Kb
0
+2Kb
0 10 20 30 40 50 60 ≥70
0 0.5 ≥1
B.
GO Biological Process
-log10(Binomial p value)
0 2 4 6 8 10 12 14
transforming growth factor beta receptor signaling pathway 14.44
apoptotic signaling pathway 10.73
response to transforming growth factor beta stimulus 8.77
cellular response to transforming growth factor beta stimulus 8.72
intrinsic apoptotic signaling pathway 7.52
regulation of translation 6.94
toll-like receptor 3 signaling pathway 6.48
TRIF-dependent toll-like receptor signaling pathway 6.42
MyD88-independent toll-like receptor signaling pathway 6.00
cellular response to decreased oxygen levels 5.86
negative regulation of transforming growth factor beta receptor signaling pathway 5.85
toll-like receptor 4 signaling pathway 5.73
response to ionizing radiation 5.72
autophagy 5.68
positive regulation of cellular catabolic process 5.65
cellular response to hypoxia 5.61
toll-like receptor 5 signaling pathway 5.60
toll-like receptor 10 signaling pathway 5.60
negative regulation of cellular catabolic process 5.55
regulation of mast cell differentiation 5.52
C.
Time
0
10
20
40
60
160
Sense
Anti-sense
Mean PRO-seq Signal (10bp bins)
0.25
0.00
−0.25
−1.0 −0.5 0.0 0.5 1.0
Distance from center of eTSS (Kb)
D.
HSF2
JUND
Deepbind scores
6 4 2 0
3 2 1 0 −1
activated
unchanged
Enhancer Activity

validation, we examined ChIP-seq data from ENCODE for untreated K562 cells, and found, as expected, that enhancer and promoter classes were both strongly enriched for acetylation of histone H3 at lysine 27 (H3K27ac), and that the promoter class was more strongly enriched for RNA polymerase and trimethylation of histone H3 at lysine 4 (H3K4me3) (Figure 1.7A). The enhancer class also showed moderate enrichment for monomethylation of histone H3 at lysine 4 (H3K4me1). These observations confirm that PRO-seq serves as an efficient single-assay approach for characterizing both transcribed enhancers and protein-coding genes in our system (Danko et al., 2015).

To better understand the role of the non-coding regulatory elements in the celastrol response, we further examined 1,479 (20%) of the 7,334 dTSS-based enhancers that were classified as differentially transcribed. We attempted to find functional enrichments for potential target genes of these differentially transcribed enhancers using the Genome Regions Enrichment of Annotations Tool (GREAT, Ver. 3.0; see Methods), which associates candidate regulatory elements with likely target genes according to distance-based rules and then tests those genes for functional enrichments (McLean et al., 2010). GREAT identified enrichments for processes relating to apoptosis, translational regulation, and responses to various environmental stresses (Figure 1.7B), in general agreement with our analysis of DE genes. We also found that our set of putative enhancers displayed an accumulation of paused polymerase after celastrol treatment (Figure 1.7C, Supplemental Figure 1.26). Although the functional significance of pausing at enhancers is unknown, this observation suggests that global shifts in pause levels at genic TSS are also reflected at enhancers.

Finally, we sought to determine which TFs influenced activity at enhancers.

Because sparse data at enhancers resulted in noisier estimates of transcriptionally engaged RNA polymerase than at genes, we focused in this case on a relatively small group of 480 enhancers that showed little activity at 0 minutes but greatly increased activity by 160 minutes (see Methods). We compared DeepBind scores for these activated enhancers with those for non-DE enhancers that had similar absolute expression levels and found six TFs whose motifs had significantly elevated scores for sequence elements in the activated enhancers: HSF1/2, JUND, FOSL2, MAFK, STAT3, and THRA (Supplemental Figure 1.27; 1.7D). Of these TFs, HSF2, was also associated with increased expression in genes, while JUND and FOSL2 are subunits of AP-1, a TF previously found to regulate cellular growth and senescence (Shaulian and Karin, 2001). Because these TFs were identified simply based on their sequence preferences, TFs with similar motifs are also potential regulators. For example, it is possible that JUN, whose expression increases over the time course and which is known to be activated by HSF1 (Sawai et al., 2013), is actually responsible for the apparent association with JUND, which does not appear to be activated. We also cannot effectively distinguish between HSF1 and HSF2 binding here. In addition, since this analysis was limited to TFs that increased transcription at enhancers, it is unsurprising that it did not identify TFs associated with decreased expression in the genic response, such as MAX. Despite these caveats, these results demonstrate that PRO-seq can be used to detect transiently activated enhancers and identify candidate TFs that may drive the enhancer response.

1.4 Discussion

This study represents the first genome-wide assessment of the immediate transcriptional effects of celastrol, including transcribed regulatory elements as well as genes, shedding light on some of the possible primary targets and mechanisms of action of this potent therapeutic compound. We find that celastrol treatment results in pervasive transcriptional down-regulation, with nearly half of expressed genes being down-regulated within 160 minutes. A much smaller group of genes, roughly 10% of those expressed, are up-regulated during the same time interval. By analyzing the sequences nearby transcription units, we were able to identify several transcription factors whose binding patterns partially explain these transcriptional responses. We also observed a clear impact from celastrol on polymerase pausing at both genes and enhancers, which is negatively correlated with changes in transcriptional activity. While there are limits to what can be learned from PRO-seq data alone, we have shown that when these data are collected at relatively high temporal resolution and analyzed together with other data for the untreated condition, they can provide valuable insights into a multifaceted, multistage cellular response to a transcriptional stimulus.

We find that celastrol treatment generally induces a similar response to heat shock, consistent with previous reports (Trott et al., 2008; Westerheide et al., 2004). Heat-shock has also been observed to induce widespread down-regulation in mammalian cells (Hieda et al., 2004; Mahat et al., 2016; Vihervaara et al., 2017). Moreover, many of the same genes that are up-regulated upon celastrol treatment are also bound by HSF1 after heat shock or participate in heat-shock pathways. In addition, sequences associated with HSF1 binding are associated with increased gene expression, according to our regression analysis. Finally, our direct comparison showed that both treatments lead to up-regulation of genes involved in the heat shock response, and down-regulation of genes involved in ribosomal formation and translation.

Together, these findings suggest that HSF1 is activated soon after celastrol treatment, whereupon it activates a large group of genes. These observations suggest that, in part, the transcriptional response to celastrol may simply be a general cellular stress response. Cellular stress responses are known to affect a broad range of cellular functions, including cell cycle arrest, transcription of molecular chaperones, activation of DNA damage repair pathways. removal of irretrievably damaged macromolecules, and apoptosis upon severe damage (de Nadal et al., 2011), and they are relevant in many diseases, including cancer (Bi et al., 2005), proteotoxic diseases (Mu et al., 2008), and autoimmune disorders (Todd et al., 2008). Previous studies have investigated cellular stress responses at the delayed transcriptional (Mahat et al., 2016; Teves and Henikoff, 2011), post-transcriptional (Gardner, 2008), translational (Shalgi et al., 2013), and post-translational levels (Golebiowski et al., 2009; Urano et al., 2000). Together with similar studies of heat shock (Mahat et al., 2016), our study helps to illuminate specific features of the early transcriptional response to stress.

Nevertheless, we observed several differences between the heat shock and celastrol responses. The most striking difference was that genes associated with the cholesterol synthesis pathway are activated by heat shock but repressed by celastrol. This observation is consistent with reports that celastrol decreases endogenous cholesterol in mice and reduces obesity (Ma et al., 2015; Liu et al., 2015; Zhang et al., 2016). Further analysis suggested that loss of SREBF1 binding may be responsible for the difference in the celastrol response. Given that SREBFs are activated via proteolytic cleavage from an intermembrane protein when sterols are scarce (Wang et al., 1994), it is possible that celastrol inhibits cholesterol synthesis by inhibiting SREBF1 cleavage and release. Alternatively, or in addition, celastrol could inhibit cholesterol synthesis by decreasing the stability of the SCAP-SREBF complex (Kuan et al., 2017; Zhang et al., 2009).

Another apparent difference between the heat-shock and celastrol responses has to do with the kinetics at genes that appear to be regulated by SRF, a transcription factor associated with HSF1/2-independent up-regulation after heat-shock. We observed an enrichment for SRF binding sites in the core promoters of genes that displayed elevated expression after 40 minutes of celastrol treatment. By contrast, previous findings for heat shock (Mahat et al., 2016) have indicated that SRF-mediated up-regulation occurs much more rapidly, as early as 2.5 minutes after treatment. Finally, our analysis suggests that the loss of binding by MYC-MAX may be responsible, in part, for the broad transcriptional repression we observe within 20 minutes of celastrol treatment. To our knowledge, MYC-MAX inhibition has not been reported to be important in the heat-shock response. The finding for celastrol is supported by previous studies showing that celastrol directly inhibits MYC-MAX functionality (Wang et al., 2015b). Since MYC-MAX is a strong transcriptional activator and is bound at over 6,000 promoters of active genes in K562, its inhibition may be an important contributing factor to widespread down-regulation after celastrol treatment (Amati et al., 1992, 1993).

A major strength of our experimental approach is that it allows us to observe transient as well as sustained transcriptional responses. For example, we found that E2F4, a transcriptional repressor, was quickly down-regulated after celastrol treatment, reducing in expression by half within 20 minutes. Despite loss of this repressor, target genes of E2F4 were increasingly down-regulated (rather than up-regulated, as expected) after 60 minutes (Lee et al., 2011). These observations suggest that the apparent increased repression activity of E2F4 during the celastrol response may have a non-transcriptional basis. Interestingly, previous studies have shown that celastrol inhibits CDK4, and that CDK4 overexpression disrupts E2F4 DNA-binding ability (Peng et al., 2010; Scimè et al., 2008). Thus, it is possible that celastrol increases the DNA binding of E2F4 to DNA, which in turn could contribute to cell-cycle arrest (Lee et al., 2011; Polager and Ginsberg, 2003).

Another advantage of our densely sampled PRO-seq time course is that it allows us to measure changes in promoter-proximal RNA polymerase pausing. We observed that pause indices increased by more than two-fold at differentially expressed genes during our time course. We also found that increase pausing was associated with decreased transcription in genes, as previously reported for heat-shock conditions (Mahat et al., 2016), although we did not observe a converse association between decreased pausing and up-regulation of genes. A possible mechanism that could contribute to this increased pausing in down-regulated genes is the celastrol-induced disruption of the MYC-MAX complex, which has been shown to recruit P-TEFb, which in turn broadly facilitates pause release (Kanazawa et al., 2003). This mechanism could in principle affect down-regulated genes only, for example, if up-regulated genes recruit P-TEFb independently of MYC-MAX (e.g., through the activity of HSF1; Lis et al. (2000)).

While it is possible that increased pausing causes decreased expression, by limiting productive elongation and therefore reducing transcription levels, an alternative possibility is that increased pausing is a consequence of "mass action"—that is, decreased transcriptional activity across many genes results in increased availability of free Pol II, some of which ends up being loaded on promoters and coming to rest at pause sites (Mahat et al., 2016). In other words, the negative correlation between pausing and expression could be explained by causality in either direction, or perhaps in both directions. Additional experiments will be needed to establish the causal basis of these correlations. In any case, our observations suggest that changes in pausing are widespread and broadly associated with transcriptional repression, and therefore may play an important role in the celastrol response.

1.5 Materials and Methods

1.5.1 Celastrol treatment

K562 cells were cultured at 37°C in RPMI media (Gibco) containing 10% FBS (Gibco), Pen Strep (Gibco) and 2 mM L-Glutamine (Gibco). Biological replicate cell cultures were prepared as follows: after thawing K562 cells and seeding a fresh culture, cells were split into two separate flasks, which would remain separated through six passages and expansions until treatment and collection for preparation of PRO-seq libraries. Cells from each expanded replicate were seeded onto six 30-mL dishes (one for each time point) at a density of 5x105 cells/mL and then incubated for an additional doubling cycle (~20 hrs). For treatments, fresh celastrol was dissolved in DMSO at a final concentration of 20 mM. Celastrol-treated samples received celastrol (Sigma) at a final concentration of 3 μM, whereas untreated (0-minute) samples received an equivalent volume of DMSO. Cells remained in culture dishes in the incubator during the time course. Time-course treatments were carried out in reverse order so that all samples would be collected at the same time (starting with 160-minute time point and ending with the untreated).

1.5.2 Cell permeablization and PRO-seq

Samples were then prepared for precision run-on reactions by subjecting cells to permeablizing conditions. Briefly, cultures were spun down and resuspended in ice cold 1xPBS. Samples were spun again and washed in 5 mL wash buffer (10 mM Tris-Cl, pH 7.5; 10 mM KCl; 150 mM sucrose; 5 mM MgCl2; 0.5 mM CaCl2; 0.5 mM DTT; 1x Protease inhibitor cocktail (Roche); 20 units RNase inhibitor (SUPERase In, Invitrogen)). Cell pellets were then resuspended in permeablization buffer (10 mM Tris-Cl, pH 7.5; 10 mM KCl; 250 mM sucrose; 5 mM $MgCl_2$; 1 EGTA; 0.05% Tween-20; 0.1% NP40; 0.5 mM DTT; 1x Protease inhibitor cocktail (Roche); 20 units RNase inhibitor (SUPERase In, Invitrogen)) and left on ice for 5 minutes. Cells were checked for penetration by trypan blue to assess permeability (~99% permeable). Cells were then washed two times in 5 mL wash buffer before being resuspended in 200 μL storage buffer (50mM Tris-Cl, pH 8.3; 40% glycerol; 5 mM $MgCl_2$; 0.1 mM EDTA; 0.5 mM DTT). A one-to-fifty dilution was prepared using 2 μL of each sample and used to take OD600 measurements. All samples were then diluted to an equal density (OD600 = 0.181) in a final volume of 110 μL of storage buffer. 5x104 pre-permeabilized S2 cells were then spiked in to each cell count-normalized sample before flash-freezing the permeablized cells and storing them at -80°C. Stored permeable cells with spike-ins were thawed on ice and each sample was subjected to the precision run-on protocol (Mahat et al. 2016a). Run-on reactions incorporated only biotinylated NTPs with no un-modified NTPs. All libraries were subjected to nine cycles of PCR amplification before size selection and gel purification.

1.5.3 Cell counting

Cells were either treated with 3μM celastrol, DMSO, or left untreated for four days. Live/dead cells were determined based on trypan blue staining. Counts were measured with an automatic cell counter (Bio-rad).

1.5.4 Read mapping

All filtered reads were removed from each fastq file, then cutadapt (v1.9.1) was run with the following options:

```
cutadapt -a TGGAATTCTCGGGTGCCAAGG -m 15
```

to remove the Illumina adapters and discard all remaining reads that were less than 15bp in length. All reads were then trimmed to 34bp in length using fastx_timmer (v0.0.13.2) to avoid biasing read mapping away from gene promoters. The trimmed reads were then aligned to the joint hg19/bdgp6 genome using the STAR aligner (v2.4.0i) (Dobin et al., 2013). Reads aligning to hg19 and bdgp6 were then separated and bigwigs were created by converting each read to a single count at its 5′ end. While human assembly hg19 was used for read mapping, we do not expect the use of the more recent hg38/GRCh38 would have an appreciable impact on our results, as the major differences between these assemblies concern alternative haplotypes and centromeric regions.

1.5.5 Detection and resolution of dTSS

dREG was run on each sample as described previously (Danko et al., 2015), producing a set of genomic intervals corresponding to predicted divergent transcription starts sites (dTSS). These initial dREG calls had fairly coarse resolution, ranging in size from several hundred to thousands of bases. We therefore applied a heuristic scanning method to identify one or more higher-resolution dTSSs within each dREG call. Briefly, this method involved sliding a window along a dREG interval and considering the relative read counts among three subintervals: a peak, a flank, and center. To identify pairs of divergent peaks, the test was applied simultaneously to each strand in a strand-specific manner, and the results were combined. Specifically, for a scan initiated at base i, the center was defined as the interval $[i, i + 110)$, the shoulder as $[i - 50, i)$, and the flank as $[i - 250, i - 150)$. Three one-sided binomial tests were performed, testing that there are fewer reads in the center than the flank, the center than the shoulder, and the flank than the shoulder. The sum of the resulting six negative log p-values (three for each strand) then became the per-base score. The best scoring window in a dREG region was taken as a dTSS. In addition, up to two other dTSSs were called if their score exceeded 20.

1.5.6 Classifications of dTSS

dTSS were classified as either enhancers or promoters based on their relative distance from the set of all TSSs annotated in GENCODE v19. To classify each dTSS, the following rules were applied: (1) if the dTSS was greater than 1 kb, and at most 1 Mb, away from the nearest annotated promoter, it was classified as an enhancer; (2) if the dTSS was within 200 bp of an annotated promoter, or it was within 1 kb of an annotated promoter and it was the closest dTSS to the promoter, it was classified as a promoter; (3) if the dTSS was between 600bp and 1 kb away from the nearest annotated promoter, and not the closest dTSS to the promoter, It was classified as an enhancer; (4) otherwise, the dTSS was classified as unknown.

1.5.7 Selection of active transcripts in K562 cells

Selection of transcripts was performed by a new program, called TuSelector. First, a list of potential transcripts was obtained from GENCODE v19. The genic regions and data were partitioned into 100bp intervals. For each gene, a set of coarse-grained overlapping transcript models was created, where for each transcript model and interval, the interval was assigned to the transcript model if, and only if, it overlapped the corresponding annotated transcribed region by more than 50% at the nucleotide level. Next, the PRO-seq read counts in each 100 bp interval were summarized by a 1 if there were reads aligned to the interval or a 0 otherwise. TuSelector computed a likelihood for each of the possible coarse-grained transcript models at a given gene, as follows:

$$\mathcal{L}(T) = \int_{\theta_t} P(\theta_t)\left(\prod_i P(X = x_i|\theta_t)^{\delta(i,T)}\right) d\theta_t \cdot \int_{\theta_u} P(\theta_u)\left(\prod_i P(X = x_i|\theta_u)^{1-\delta(i,T)}\right) d\theta_u$$

where T is the transcript model, the products range across genomic intervals i, x_i is the summary of the data in interval i (0 or 1), $\delta(i, T)$ is an indicator function that takes value 1 when interval i is included in T and 0 otherwise, X is a Bernoulli random variable, and θ_t and θ_u are the parameters for this random variable in the transcribed and untranscribed states, respectively. $P(\theta_t)$ is assumed to be uniform over the interval $(0.3, 1)$, and $P(\theta_u)$ is assumed to be uniform over the interval $(0.01, 0.03)$. In practice, we discretized θ_u into segments of size 0.01 and θ_t into segments of size 0.05, and approximated the integrals with finite sums. Finally, in addition to the annotated transcripts, we considered a competing model representing a completely untranscribed gene.

TuSelector was run separately for each replicate and time point, and potentially produced discordant transcript calls across these runs. Therefore, we selected at most one "consensus" transcript model per gene for use in further analysis, as follows. To be considered a consensus call, TuSelector had to identify the same transcript model at least 80% of the time with at least 50% of replicate pairs both having the same transcript call. Two transcript models were considered "the same"if their endpoints differed by less than 500bp. If no transcript model met these criteria, the gene was not considered in further analysis.

1.5.8 Estimating expression and detecting differentially expressed genes

For all active, protein-coding transcripts, reads were taken from up to the first 16 kb of the gene, minus the first 500bp to avoid an influence from promoter-proximal pausing. This strategy allowed us to focus on the most recent transcription at each time point and avoid averaging over time. The maximum interval of 16 kb was based on a minimum interval between time points of 10 minutes and an average polymerase transcription rate of ~2 kb/min, minus a few kilobases of "padding". Any genes that were shorter than 700 bp were removed from the analysis. A size factor for each sample was obtained by taking the number of spike in reads per sample divided by the median number of spike in reads per sample. To estimate expression of transcriptional enhancers, reads were taken from 310 bases (assuming a 110-base spacing between dTSS as reported by Core et al. plus 100bp to either side) centered on the dTSS. Both sets of read counts were fed jointly into DESeq2, and enhancers and genes were subsequently separated for further analysis. An enhancer or gene was called as DE with an FDR ≤ 0.01 using a likelihood ratio test.

1.5.9 Clustering differentially expressed genes

Gene expression log-fold changes were computed relative to the untreated (zero-minute) time-point using the DESeq-based estimates of absolute expression (rlog values). All DE genes were then fed into the autoregressive clustering program EMMIX-WIRE using default settings. Likelihood values for between two and ten clusters were computed. We selected four clusters as a value at which the increases in likelihood with the number of clusters began to decline. To check for the robustness of our selection we repeated our analyses with five and seven clusters and found that they were not highly sensitive to the cluster number.

1.5.10 Computing functional enrichment for gene clusters

Reactome (v52) was used to assign genes to functional categories. Genes that were not annotated in Reactome were removed. The background set for all enrichments was the set of DE genes present in Reactome. Odds ratios were computed per cluster (c) and pathway (p) as:

$$OR = \frac{X_{c,p}/X_{c,\neg p}}{X_{\neg c,p}/X_{\neg c,\neg p}}$$

where X represents a count and c and $\neg c$ denote the sets of genes in, and not in, cluster c, respectively, and, similarly, p and $\neg p$ denote the sets of genes in, and not in, pathway p. An empirical null distribution of odds ratios was computed by randomly shuffling the gene assignments to pathways 100,000 times. P-values were then computed from this distribution and the Benjamini-Hochberg procedure was applied to estimate false discovery rates (FDRs).

1.5.11 Characterizing genic regulation

ChIP-seq data was downloaded from the ENCODE website (`https://www.encodeproject.org`) in narrowPeak format (optimal idr) on Sep. 30th, 2016. Scores for each gene-TF pair were computed by taking the peaks with the maximum signal that intersected [-200,+500] around the promoter. DeepBind v0.11 was run over [-200,+500] around the promoter with standard settings using all non-deprecated motifs for DNA binding proteins (Alipanahi et al., 2015). DeepBind and ChIP-seq scores were then standardized to control for differences in range. To analyze TFs that may be involved in different regulatory patterns we linearly regressed genic expression (as estimated by DESeq2) against scores from ChIP-seq or DeepBind with time point specific coefficients for each TF and a time agnostic, gene-specific coefficient to capture the fixed effect of unmodeled regulation (Alipanahi et al., 2015; Neph et al., 2012). In this framework, the expected expression of a given gene *i* at time *j* is expressed as:

$$Y_{i,j} = \beta_i + \beta_j + \sum_k \beta_{ijk} x_{ik}$$

where β_i is the gene-specific expression bias term, β_j is a time-specific bias term, and $\beta_i jk$ is the coefficient for the time-point specific effect of a TF *k*. Standard deviations for each coefficient were estimated via 1000 bootstraps. Finally, the list of TFs was filtered to keep those with FDR ≥ 0.01 in at least one time point and with a maximal change between any two coefficients in the 90th percentile. This procedure selected for TFs having an effect that was both statistically significant and of large magnitude. A set of per TF F-statistics was also calculated and are available as supplemental tables (Supplemental Table 1.1,1.2).

1.5.12 Identification and analysis of genic pause peaks

To locate pause peaks, we scanned each active transcript (see above) greater than 1 kb in length in the region of the annotated TSS ([TSS-200,TSS+200]), taking the number of reads in a 50-bp sliding window, with a sliding increment of 5 bp. The window with the largest number of reads in the untreated condition (0-minute time point) was designated as the pause peak. To compute a log2 pause index (*LPI*), we subtracted the DESeq-estimated log2 read count (the "rlog" value) for the gene body from the equivalent DESeq-estimated log2 read count at the peak. Furthermore, to compute changes in this value over time, we subtracted the *LPI* for the zero time-point from the *LPI* for each subsequent time point; that is, the change in *LPI* at time *t*, denoted ΔLPI_t, was given by $\Delta LPI_t = LPI_t – LPI_0$. Notice that normalizing changes in the pause peak by changes in the gene body in this way only increases if the number of reads in the peak increases by more than the number of reads in the gene body.

1.5.13 Analysis of heat shock data

PRO-seq heat shock gene expression values were computed in an identical manner to the celastrol data with the exception of the computation of size factors, which obtained directly from a previous analysis (Vihervaara et al., 2013) Gene transcripts were the same as those used for the celastrol data. P-values were (re-)computed using the Wald test instead of the LRT for both the celastrol and heat-shock data to allow for a single time-point analysis. Differences in the distributions of gene expression within a pathway between HS and celastrol responses were evaluated using the Kolmogorov-Smirnov test.

1.5.14 Estimating expression in enhancers

To prevent contamination from genic transcription, all dTSS previously annotated as enhancers were extended by 1 kb to either side and removed if any part of the extended enhancer was within 5 kb of a gene body. DESeq was used to estimate the transcription level in the enhancer peaks, tails, and the whole enhancer body. The enhancer peak was defined as ±250 bp from the center of the enhancer, the tail was ±400 to ±1000 bp from the center, and the whole enhancer was 0 to ±1000 bp from the center of the enhancer. Read counts were summed from both strands for each region (i.e., peak = plus strand [0,+250]+

minus strand[-250,0]), and then DESeq2 was used to estimate fold changes. Enhancers were called as strongly activated if they were differentially expressed with FDR ≤ 0.01, rlog(expression at 0 min) ≤ 1, and were in the 90th percentile for fold change between 0 and 160 minutes. To get the same number of similarly expressed non-differentially expressed enhancers, we performed rejection sampling on enhancers that were differentially expressed with FDR ¿ 0.5 using their average expression values between 0 and 160 minutes and probabilities calculated from a kernelized histogram of the activated enhancer's expression at 160 minutes.

1.5.15 Data Access

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; `http://www.ncbi.nlm.nih.gov/geo/)` under accession number GSE96869.

1.6 Supplemental Figures

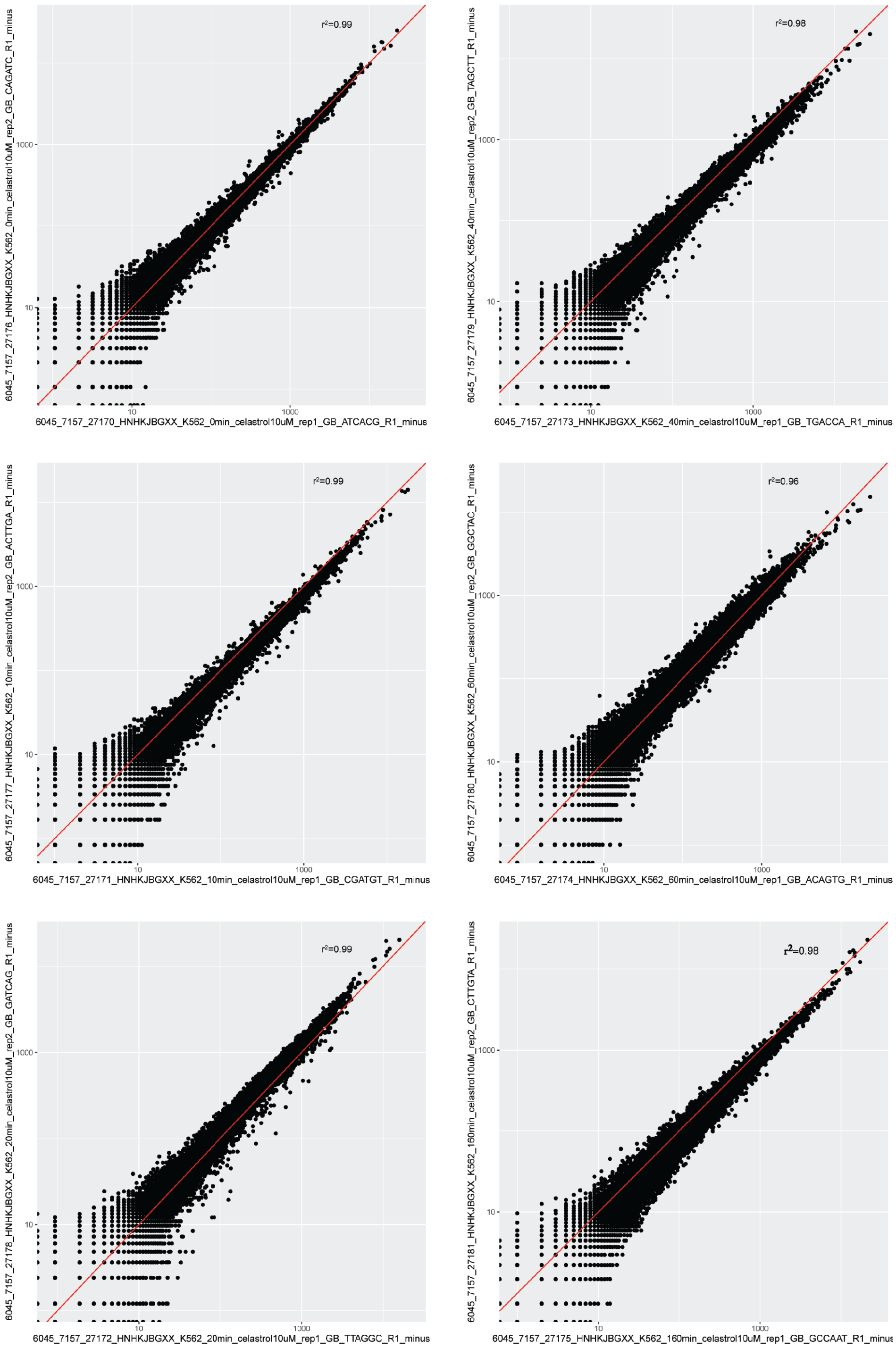


Figure 1.8: **Concordance in read-counts between replicates in gene bodies on the negative strand.**

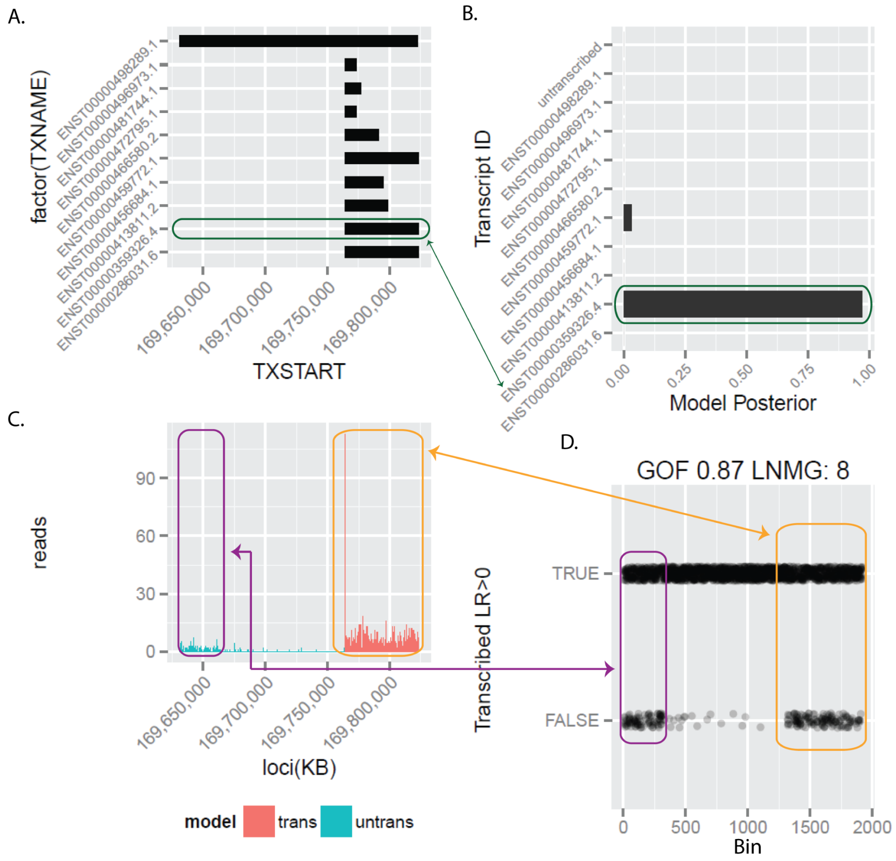


Figure 1.9: **Example of tuSelector call in the presence of multiple GENCODE transcripts.** (A) Illustration of the coordinates of each uniquely identifiable candidate transcript model (obtained from GENCODE annotations) for C1orf122. All models differ in state (transcribed vs. untranscribed) in at least one bin. (B) The posterior probability of each transcript model as computed by tuSelector. (C) The raw PRO-seq data labeled with the state calls (transcribed/untranscribed) for the most likely transcript model. Notice in this case that there is a weak indication of transcription near the start of the longest transcript model, but the absence of transcription in the middle region, and the difference in read depth compared with the stronger evidence at the right (in red), suggest that a shorter transcript model is much more likely overall. (D) Some performance metrics to compute how well the data matches the most likely annotated transcript. GOF is the fraction of bins at which an independent prediction of transcription (not considering other bins) agrees with the corresponding state in the most likely transcript model. LNMG is the longest non-matching group of consecutive bins whose individual state calls do not agree with the annotated state in the most likely model. Each dot matches a 50-bp bin of reads. If labeled “true,” the independent state calls agree with the state in the most likely model, and if labeled “false,” they do not agree.

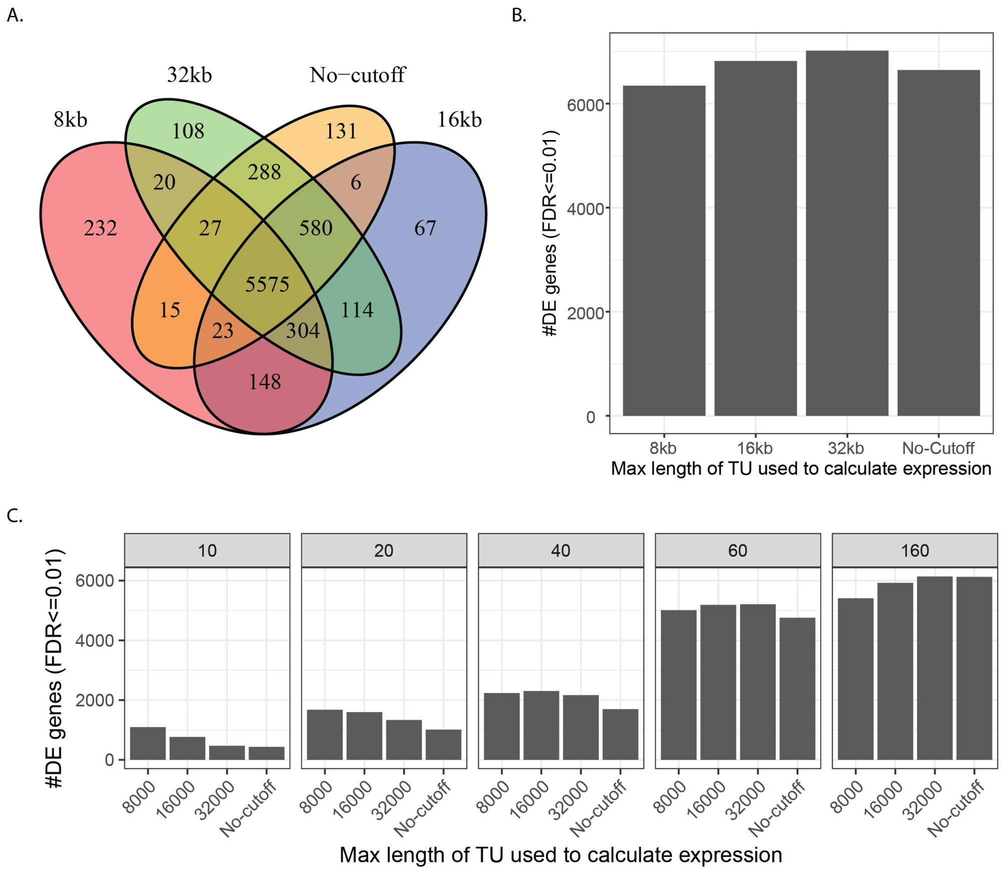


Figure 1.10: **Validation of choice of 16 Kb interval for estimating transcription levels of genes.** Here we compare 16 Kb intervals with alternative intervals of 8 Kb and 32 Kb, in all cases starting 500bp after the TSS (to avoid pausing artifacts). For example, the 8Kb label indicates that reads were counted in the interval $[TSS + 500, TSS + 8500)$. (A) Overlap between gene sets called as differentially expressed (DE) using the likelihood ratio test (LRT) at FDR ≤ 0.01 for different interval lengths. (B) Total numbers of genes called as DE using the likelihood ratio test (FDR ≤ 0.01) for different interval lengths. (C) Numbers of genes called as DE per interval length at each individual time-point using the Wald test, since our LRT aggregates across time-points (FDR ≤ 0.01). The results are similar across interval lengths, but the shorter lengths result in somewhat decreased numbers of DE genes at later time points due to reduced statistical power deriving from consideration of fewer sequencing reads. At the same time, the shorter lengths result in slightly improved sensitivity for differential expression at 10 minutes. Overall, the 16 Kb length cutoff appears to strike a good balance between sensitivity for immediate changes of large effect and sensitivity for delayed changes of modest effect.

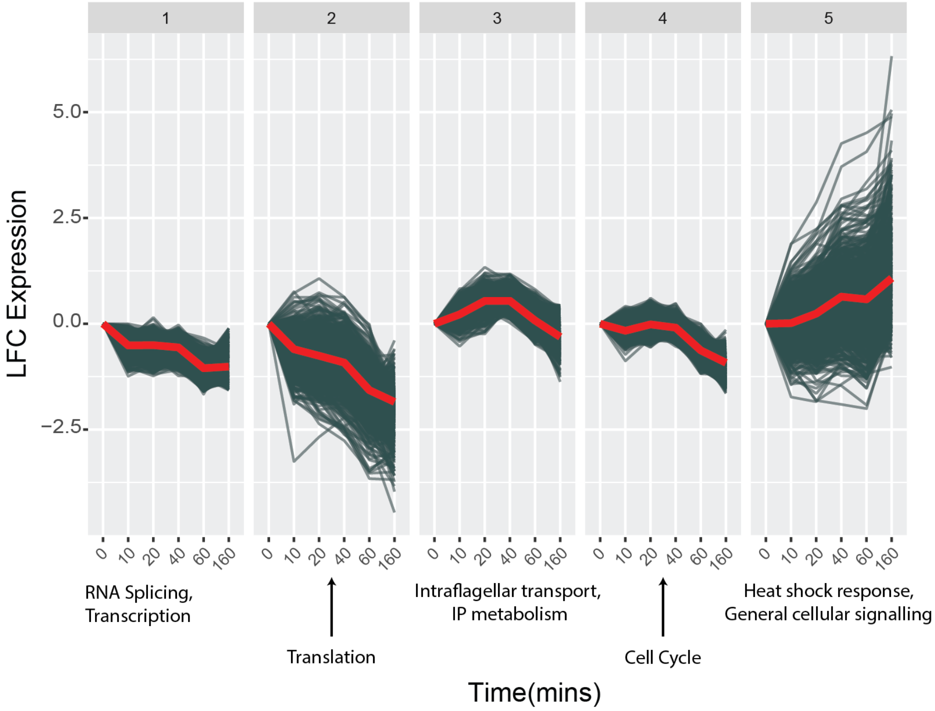


Figure 1.11: **Clustering of DE genes into five clusters and summary of enriched cluster-specific terms.**

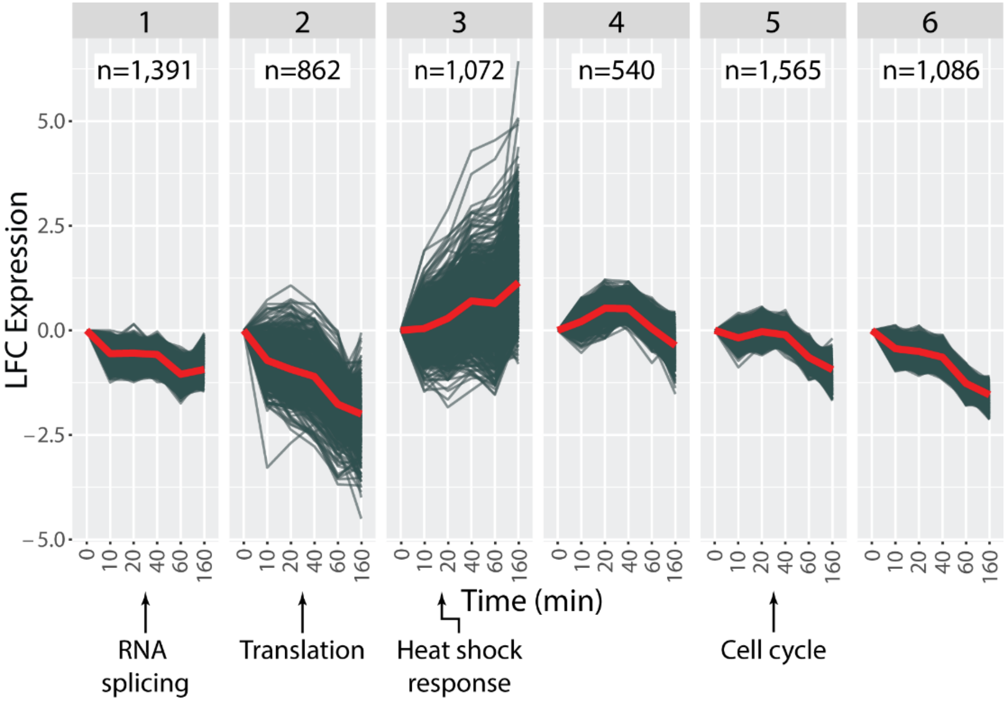


Figure 1.12: **Clustering of DE genes into six clusters and summary of enriched cluster-specific terms.**

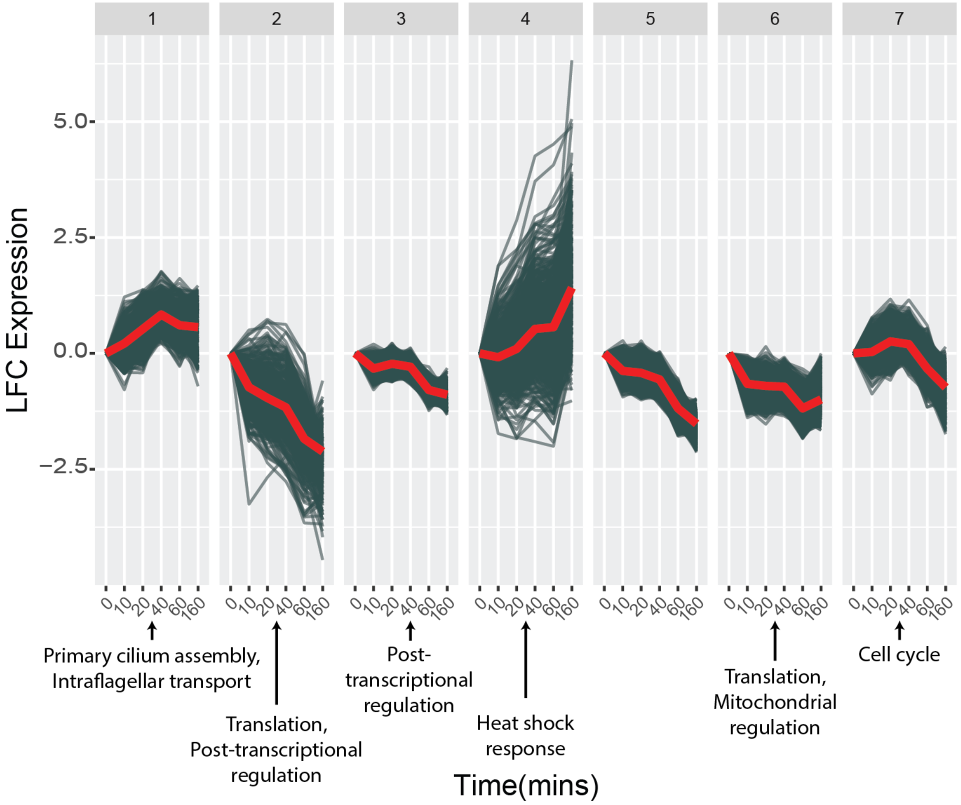


Figure 1.13: **Clustering of DE genes into seven clusters and summary of enriched cluster-specific terms.**

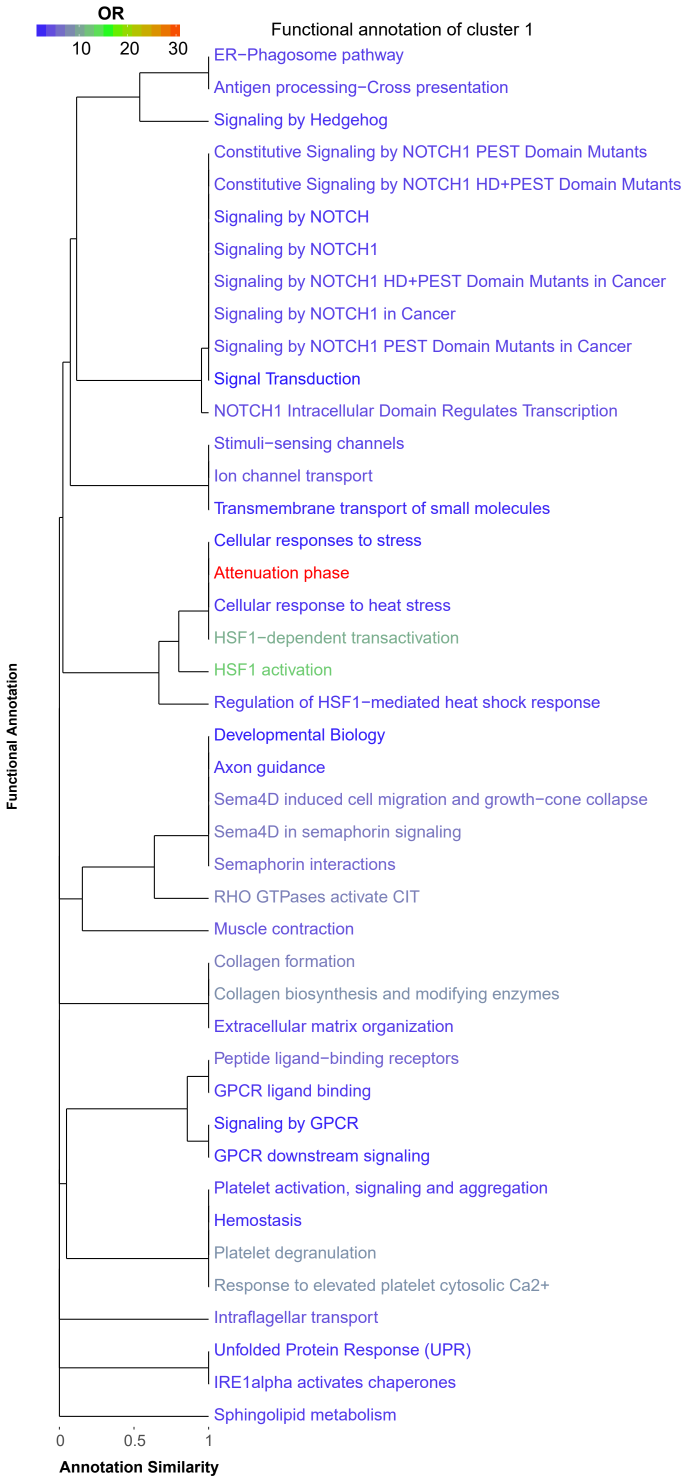


Figure 1.14: **Full set of Reactome terms enriched in cluster #1 with respect to the other clusters (FDR ≤ 0.05).** Annotation similarity indicates what fraction of genes (based upon the term associated with less genes) are shared between two terms.

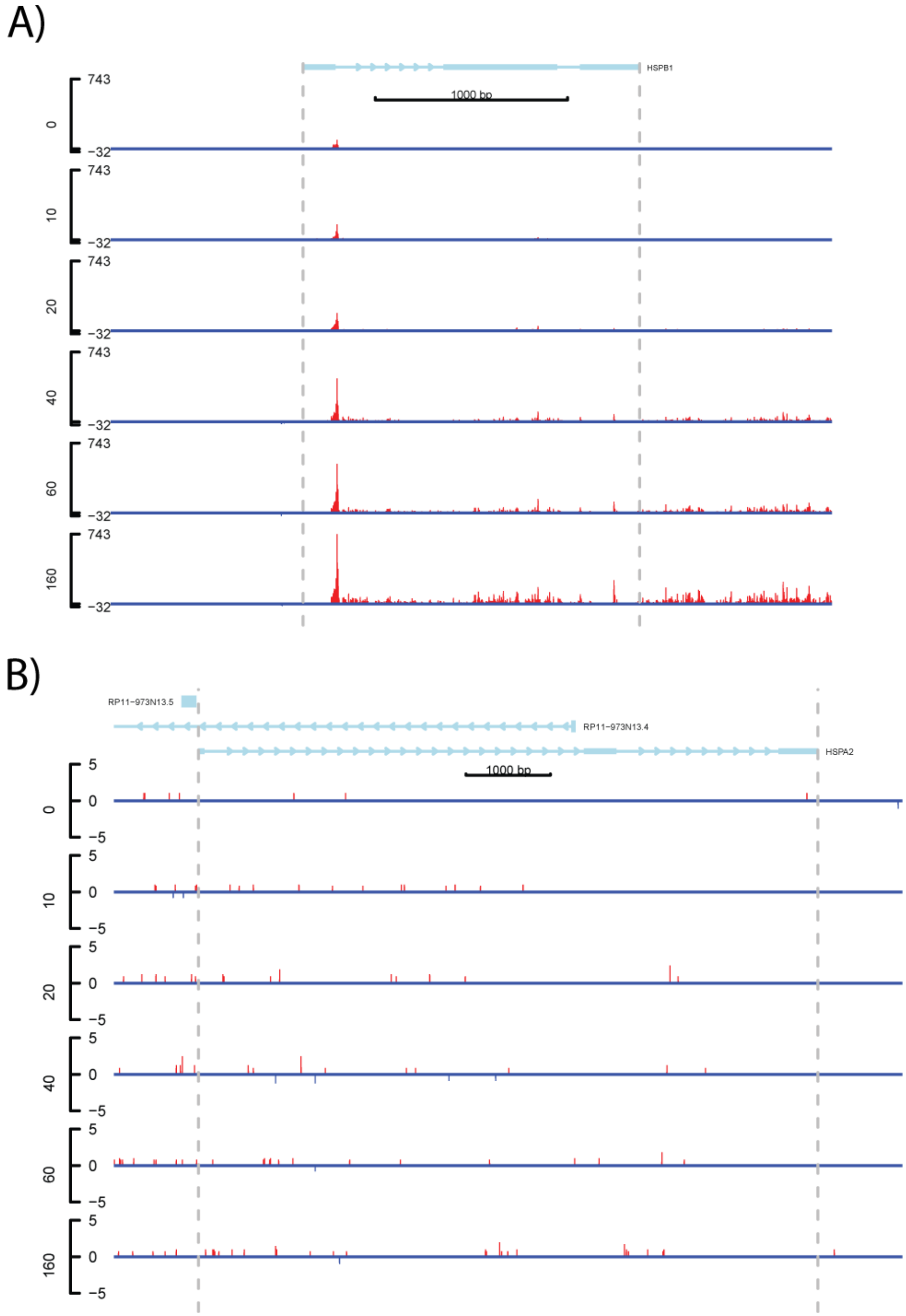


Figure 1.15: **Expression of HSPB1 and HSPA2, key genes in heat-shock induced translational repression.** (A) Expression of HSPB1 with each library normalized by size factor and replicates for each time point added together. (B) Expression of HSPA1 with each library normalized by size factor and replicates for each time point added together.

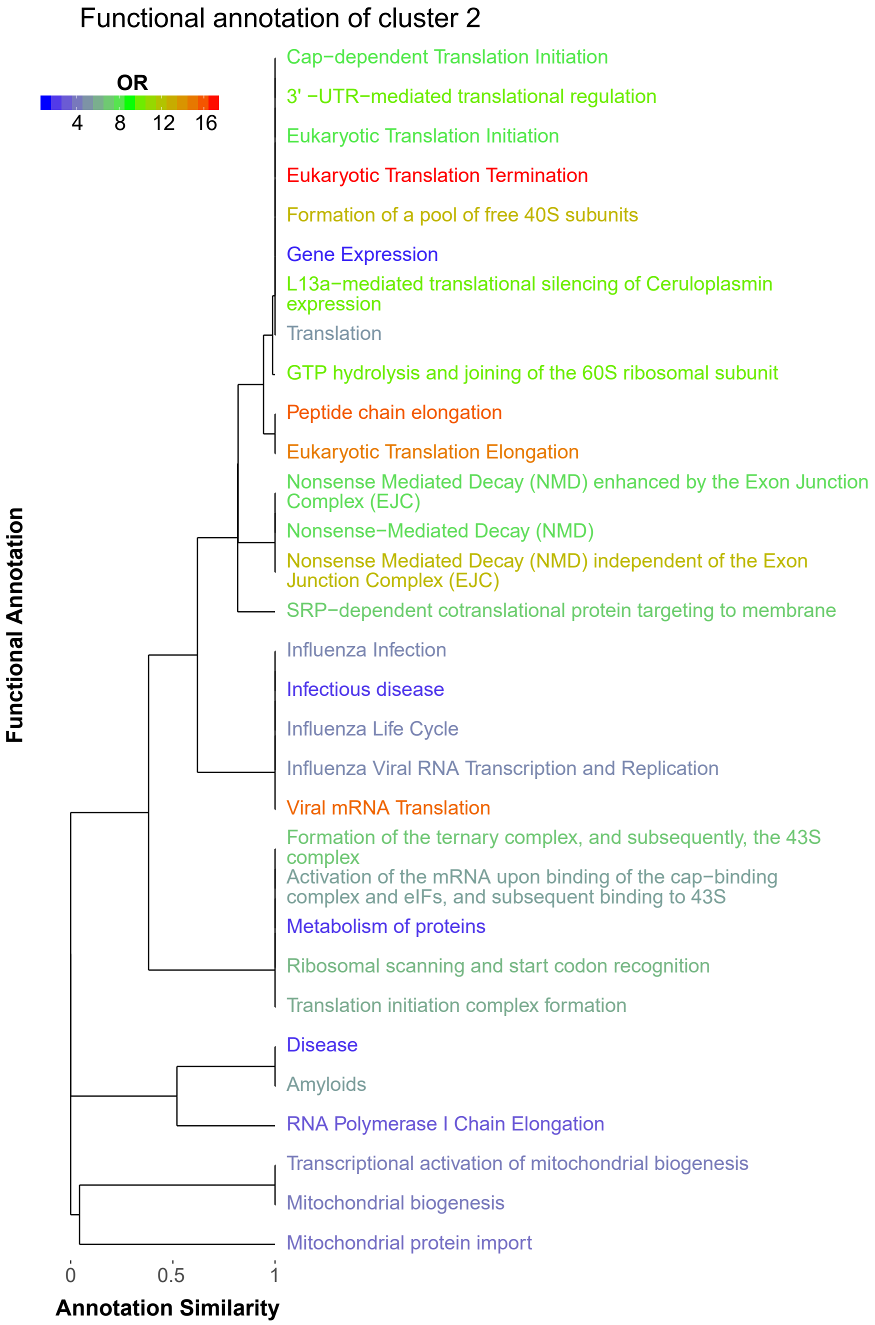


Figure 1.16: **Full set of Reactome terms enriched in cluster #2 with respect to the other clusters (FDR ≤ 0.05).** Annotation similarity indicates what fraction of genes (based upon the term associated with less genes) are shared between two terms.

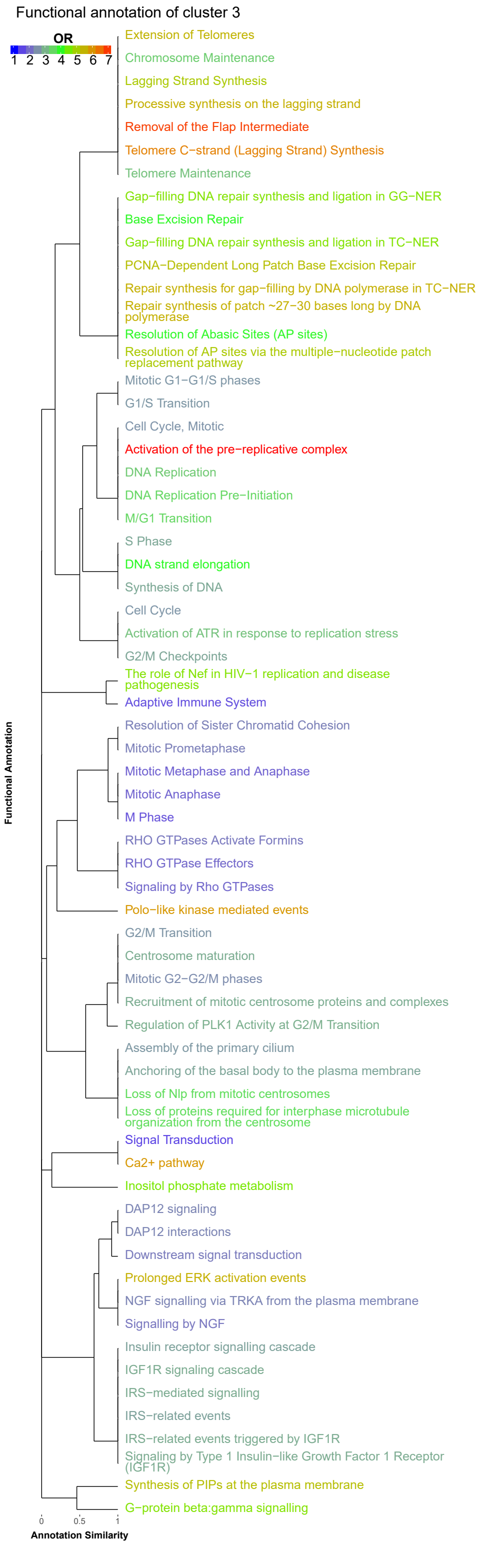


Figure 1.17: **Full set of Reactome terms enriched in cluster** #3 **with respect to the other clusters (FDR ≤ 0.05).** Annotation similarity indicates what fraction of genes (based upon the term associated with less genes) are shared between two terms.

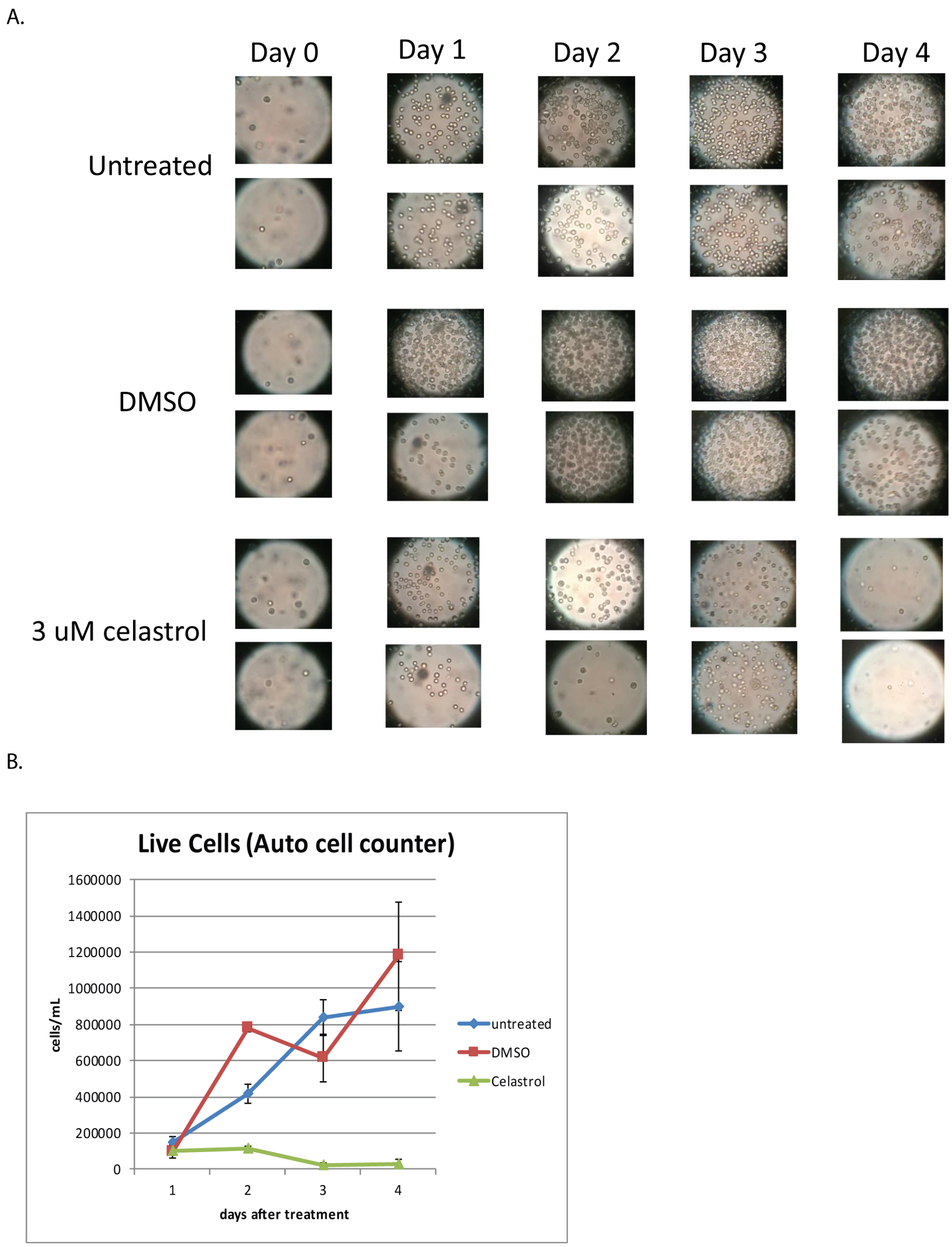


Figure 1.18: **Celastrol inhibits cell proliferation.** (A) Images of cell cultures over 4 days, two replicates of three conditions: untreated, DMSO, and 3μM celastrol. (B) Number of live cells in each cell culture as counted by an automatic cell counter (TC20, Bio-Rad).

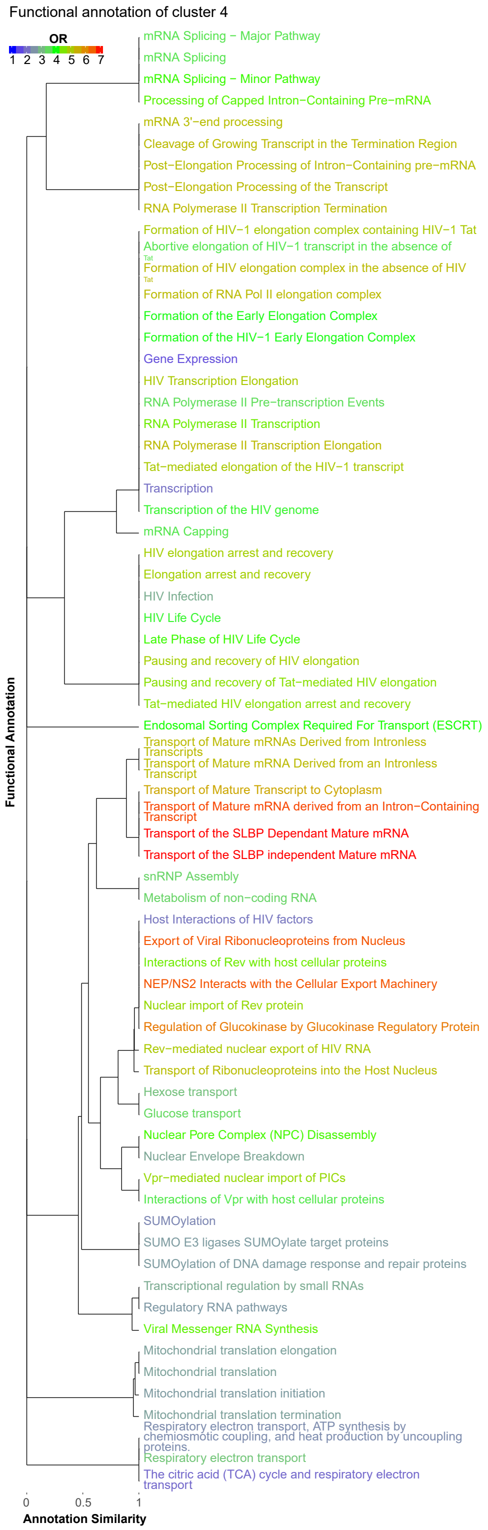


Figure 1.19: **Full set of Reactome terms enriched in cluster #4 with respect to the other clusters (FDR $\leq$ 0.05).** Annotation similarity indicates what fraction of genes (based upon the term associated with less genes) are shared between two terms.

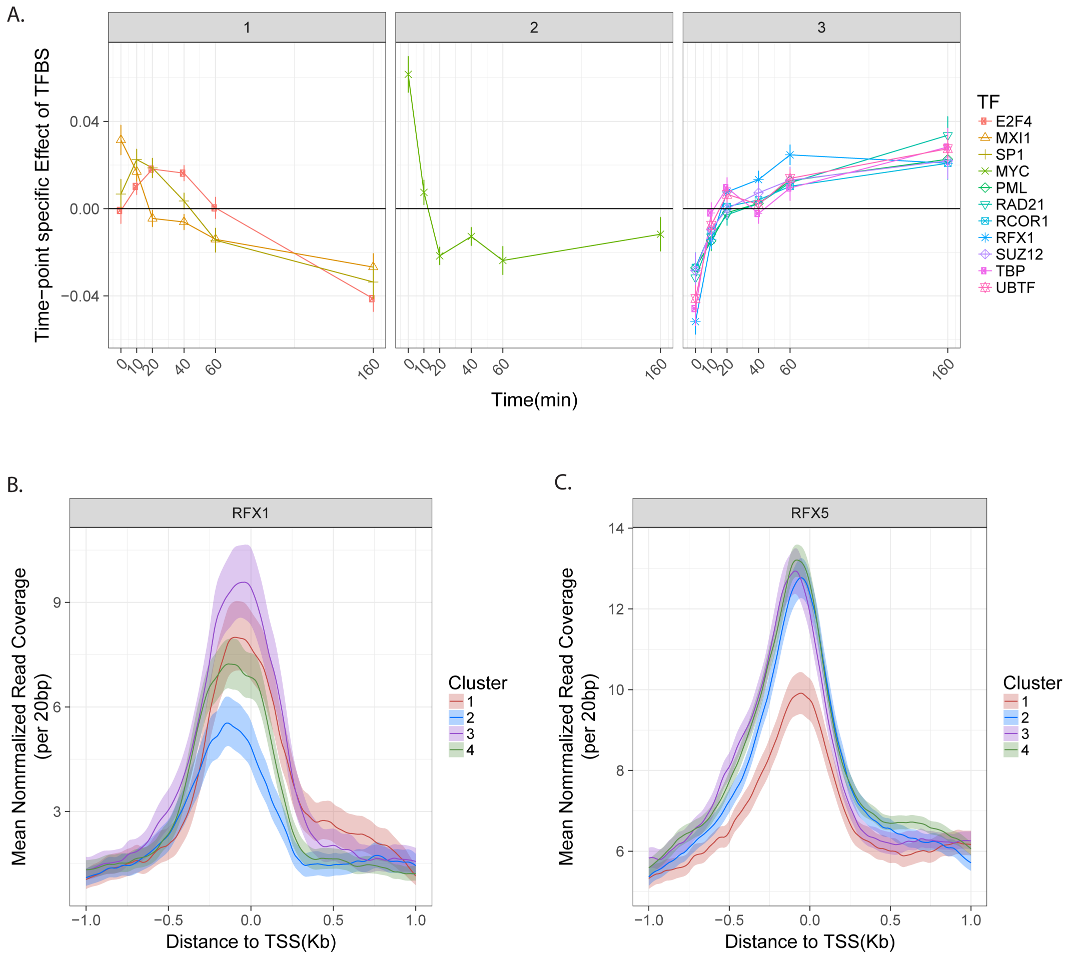


Figure 1.20: **TFBS regression model with ChIP-seq data.** (A) Weights from the regression model to predict gene expression from ChIP-seq peak scores in K562 cells. A positive weight for a TF at given timepoint means that genes at which that TF is predicted to bind in the promoter region showed increased expression relative to those without binding by that TF. Negative weights mean the opposite. Time-point specific TF coefficients explain ∽ 15% of the residual variance not explained by gene-specific or time-point-specific terms. (B) Pre-celastrol treatment ChIP-seq signal for RFX1 grouped by clustering based on expression profiles. Each line represents an average over all genes in the cluster in the region of the TSS, with lighter-colored bands representing 95% confidence intervals obtained by bootstrap sampling. (C) Same as (B) but for RFX5.

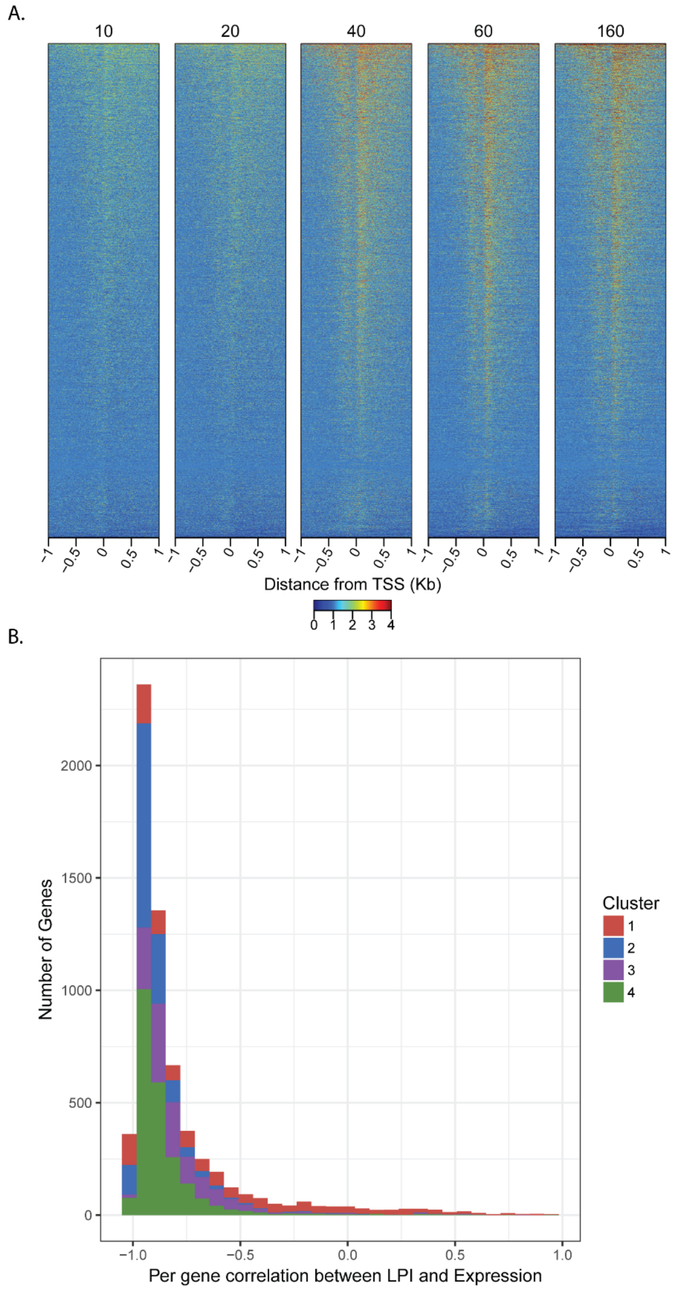


Figure 1.21: **Increased pausing at promoters is pervasive and broadly anti-correlated with gene expression.** (A) Heatmap of fold-change in read counts near promoters of all protein coding genes relative to the 0-minute time point. Numbers at top indicate minutes after celastrol treatment. (B) Histogram of per-gene Pearson correlations between log pause index (LPI) and gene expression. Cluster labeling is from the clustering of gene expression profiles in Figure 3. For most genes, pausing and expression are anti-correlated, but cluster #1 has a long tail of genes with correlations near, or greater than 0.

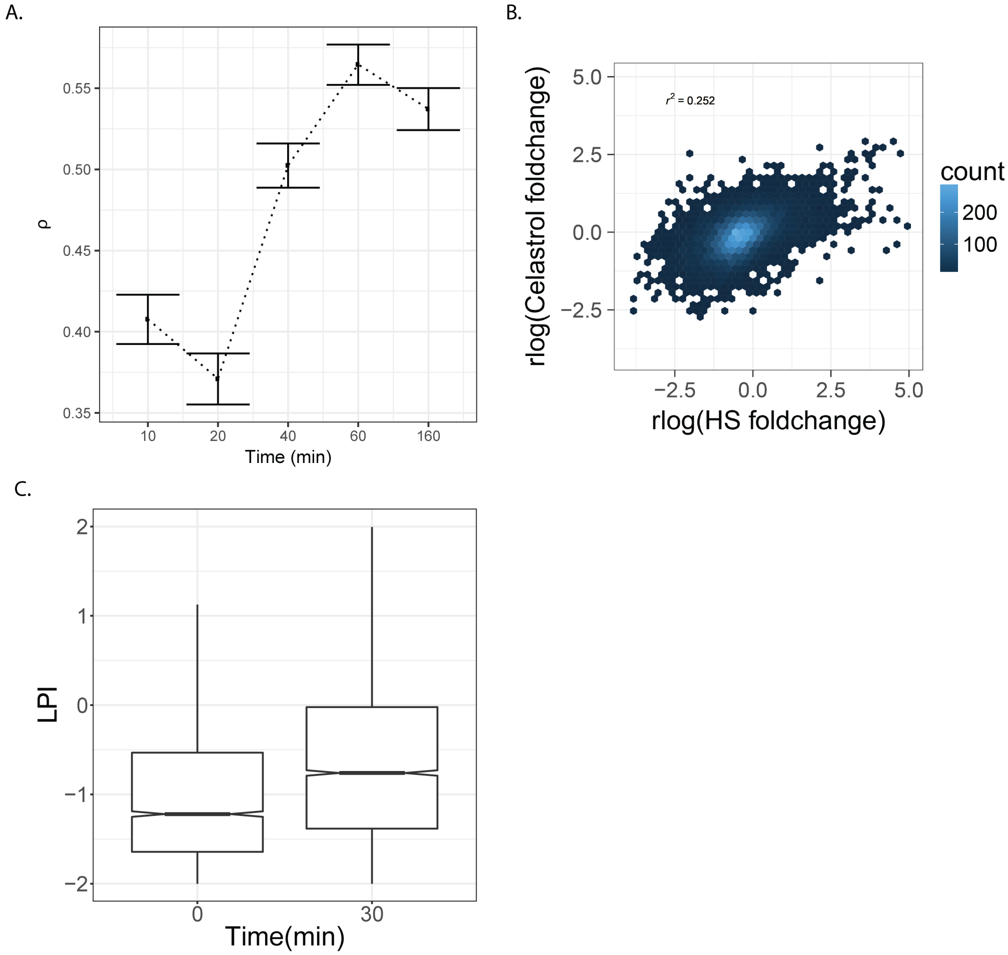
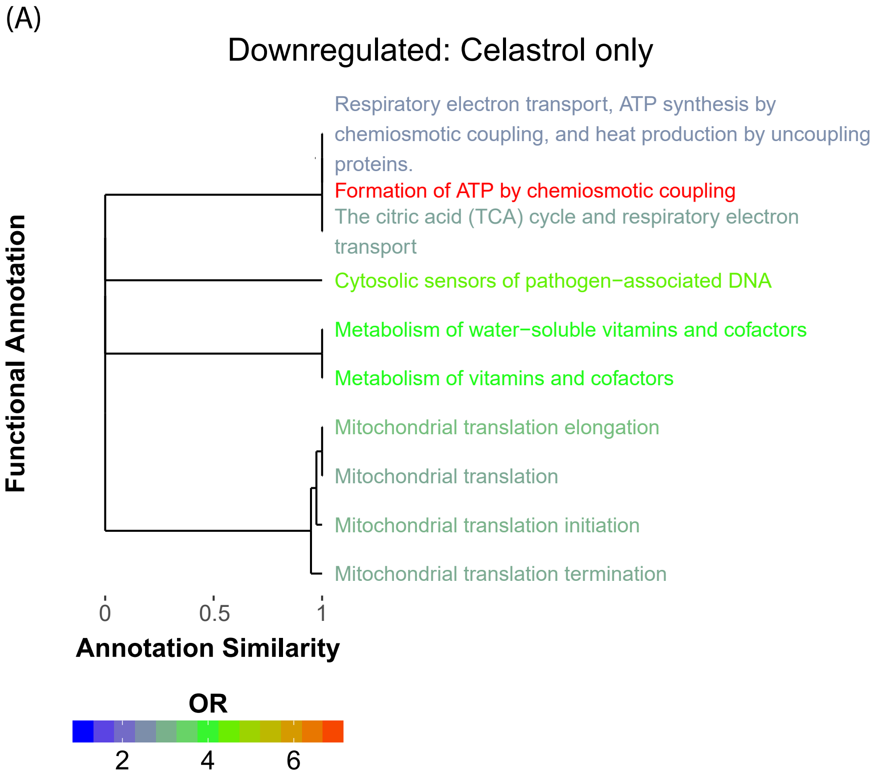
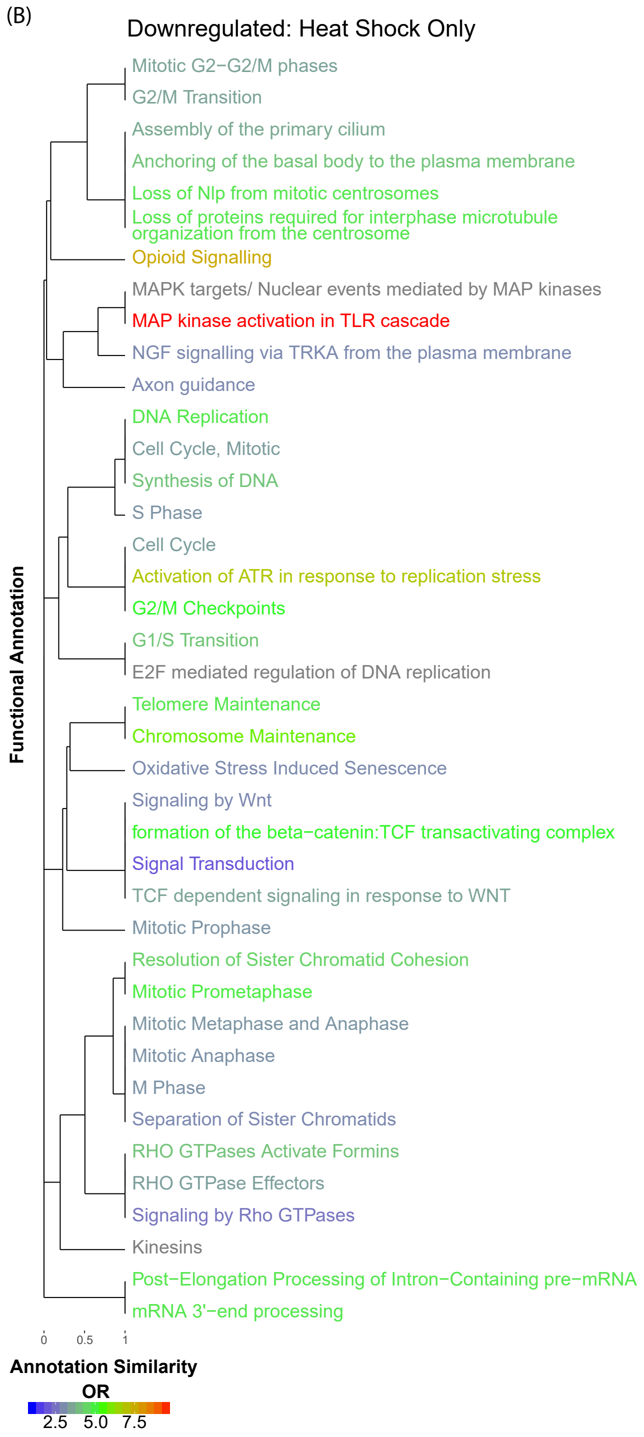


Figure 1.22: **Heat shock induces a similar, but more pronounced, response than celastrol.** (A) Pearson correlation of gene expression after 30 minutes of heat shock (Vivevaara et al., in press) with gene expression at each time-point after celastrol treatment. Error bars represent 95% CIs of the mean computed using Fisher's method for computing errors of Pearson correlation coefficients. (B) Distribution of fold changes in gene expression after 30 minutes of heat shock vs. after 40 minutes of celastrol treatment. (C) Distribution of log fold pause indices for all active genes at each time point during heat shock.


(A)
Downregulated: Celastrol only
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
Formation of ATP by chemiosmotic coupling
The citric acid (TCA) cycle and respiratory electron transport
Cytosolic sensors of pathogen−associated DNA
Metabolism of water−soluble vitamins and cofactors
Metabolism of vitamins and cofactors
Mitochondrial translation elongation
Mitochondrial translation
Mitochondrial translation initiation
Mitochondrial translation termination
Functional Annotation
0
0.5
1
Annotation Similarity
OR
2
4
6


(B)
Downregulated: Heat Shock Only
Mitotic G2−G2/M phases
G2/M Transition
Assembly of the primary cilium
Anchoring of the basal body to the plasma membrane
Loss of Nlp from mitotic centrosomes
Loss of proteins required for interphase microtubule organization from the centrosome
Opioid Signalling
MAPK targets/ Nuclear events mediated by MAP kinases
MAP kinase activation in TLR cascade
NGF signalling via TRKA from the plasma membrane
Axon guidance
DNA Replication
Cell Cycle, Mitotic
Synthesis of DNA
S Phase
Cell Cycle
Activation of ATR in response to replication stress
G2/M Checkpoints
G1/S Transition
E2F mediated regulation of DNA replication
Telomere Maintenance
Chromosome Maintenance
Oxidative Stress Induced Senescence
Signaling by Wnt
formation of the beta−catenin:TCF transactivating complex
Signal Transduction
TCF dependent signaling in response to WNT
Mitotic Prophase
Resolution of Sister Chromatid Cohesion
Mitotic Prometaphase
Mitotic Metaphase and Anaphase
Mitotic Anaphase
M Phase
Separation of Sister Chromatids
RHO GTPases Activate Formins
RHO GTPase Effectors
Signaling by Rho GTPases
Kinesins
Post−Elongation Processing of Intron−Containing pre−mRNA
mRNA 3'−end processing
Functional Annotation
0
0.5
1
Annotation Similarity
OR
2.5
5.0
7.5

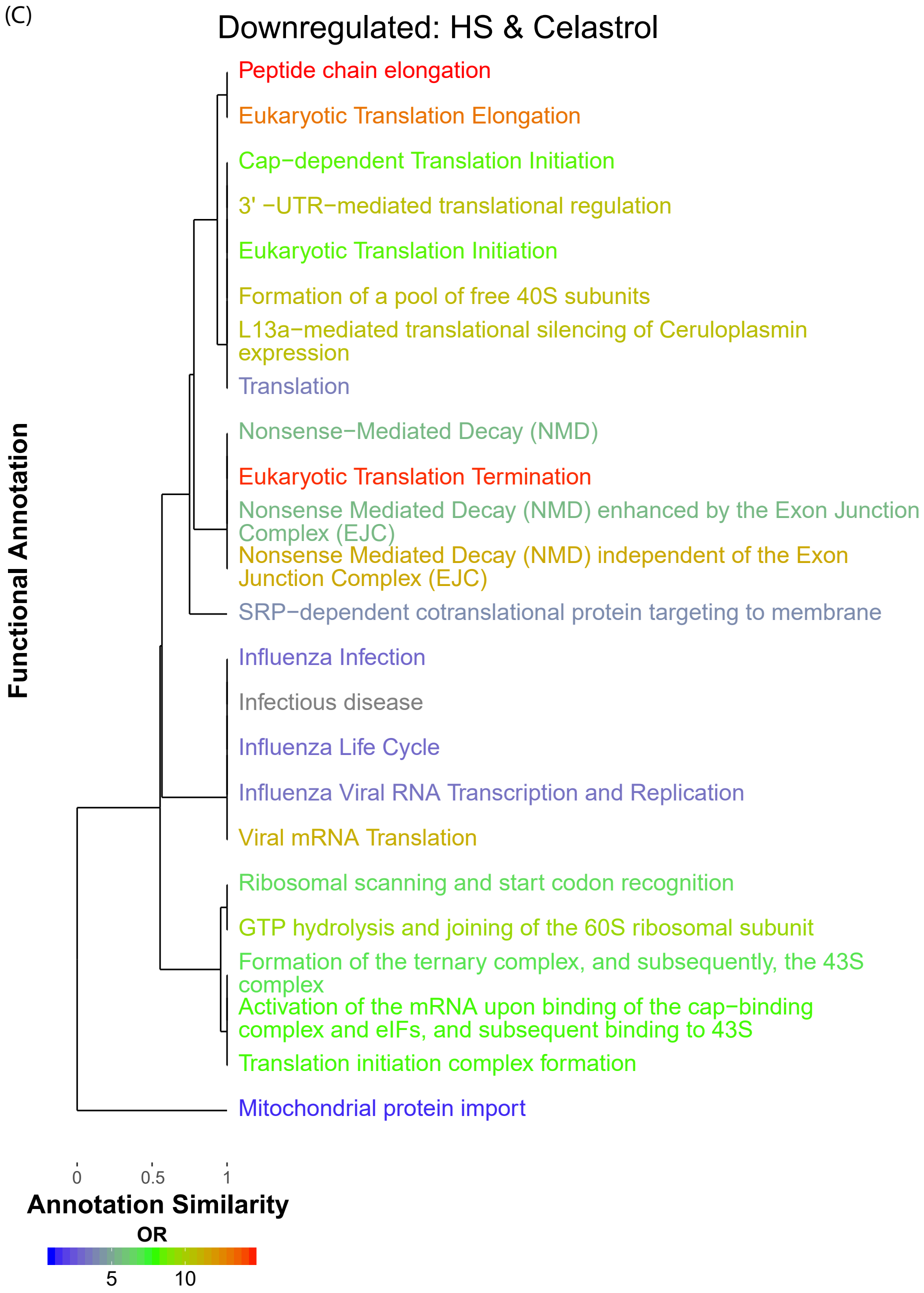


Figure 1.23: **The shared repressive celastrol and heatshock response is dominated by translation and nonsense mediated decay.** (A) Enrichment of celastrol (40 min.) specific repressed genes relative to all repressed genes (FDR ≤ 0.05). (B) Enrichment of heat shock (30 min.) specific repressed genes relative to all repressed genes (FDR ≤ 0.05). (C) Enrichment of genes repressed in both the heat shock and celastrol treatments relative to all repressed genes (FDR ≤ 0.05).

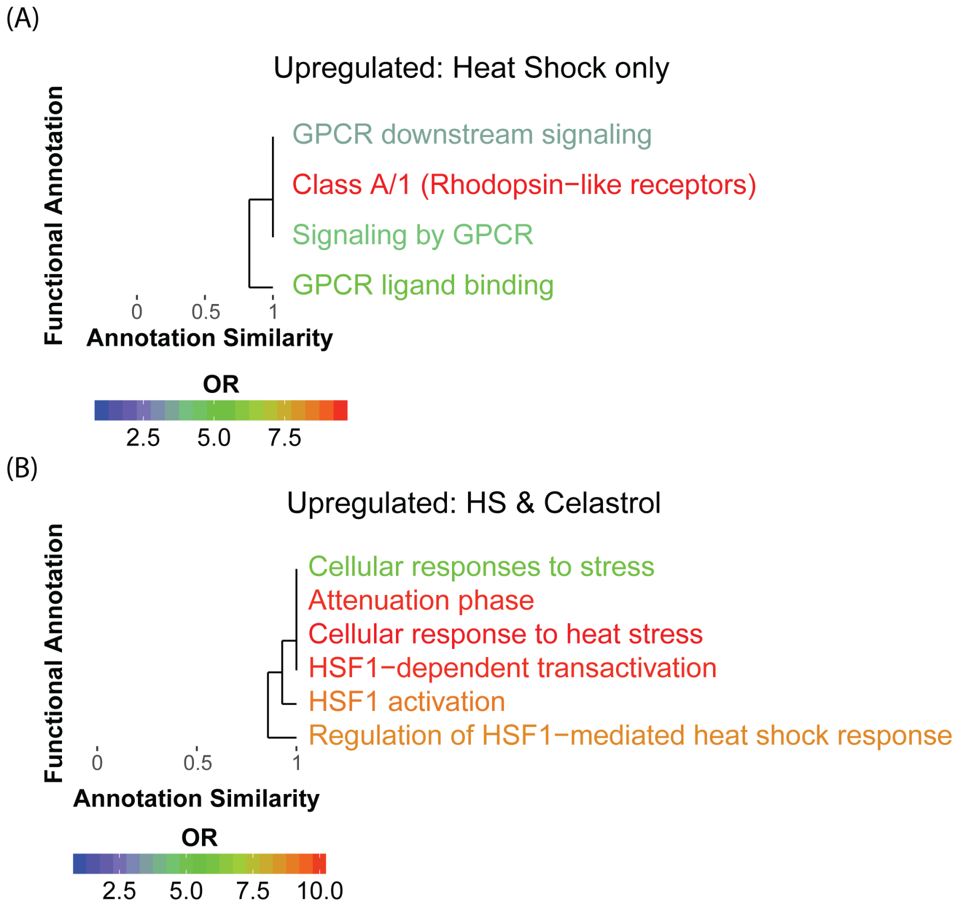


Figure 1.24: **The shared celastrol and heat-shock response for up-regulated genes is dominated by the HSF response.** (A) Enrichment of heat shock (30 min.)-specific activated genes relative to all activated genes (FDR ≤ 0.05). (B) Enrichment of genes repressed in both the heat shock and celastrol responses relative to all repressed genes (FDR ≤ 0.05).

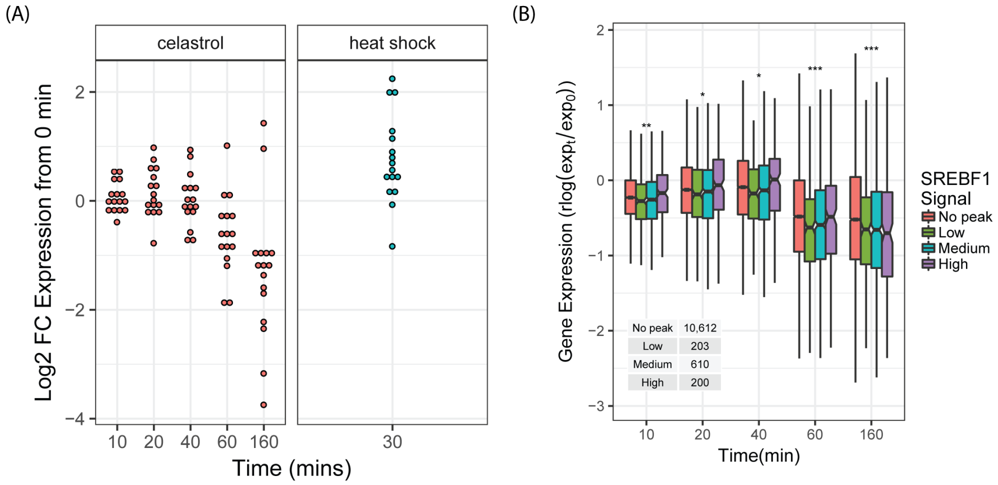


Figure 1.25: **Heat shock and celastrol have opposite effects cholesterol biosynthesis associated genes.** (A) Log2 fold-changes in expression in the celastrol (left) and heat shock (right) responses of genes that are involved in cholesterol biosynthesis and that are differentially expressed in at least one sample (FDR ≤ 0.01). (B) Log2 fold-change expression of genes stratified by status of ChIP-seq signal for SREBF1 within 500bp of the promoter, based on ChIP-seq data for untreated K562 cells from ENCODE. Low: <20th percentile; Medium: 20th–80th percentile; High: > 80th percentile in peak intensity.

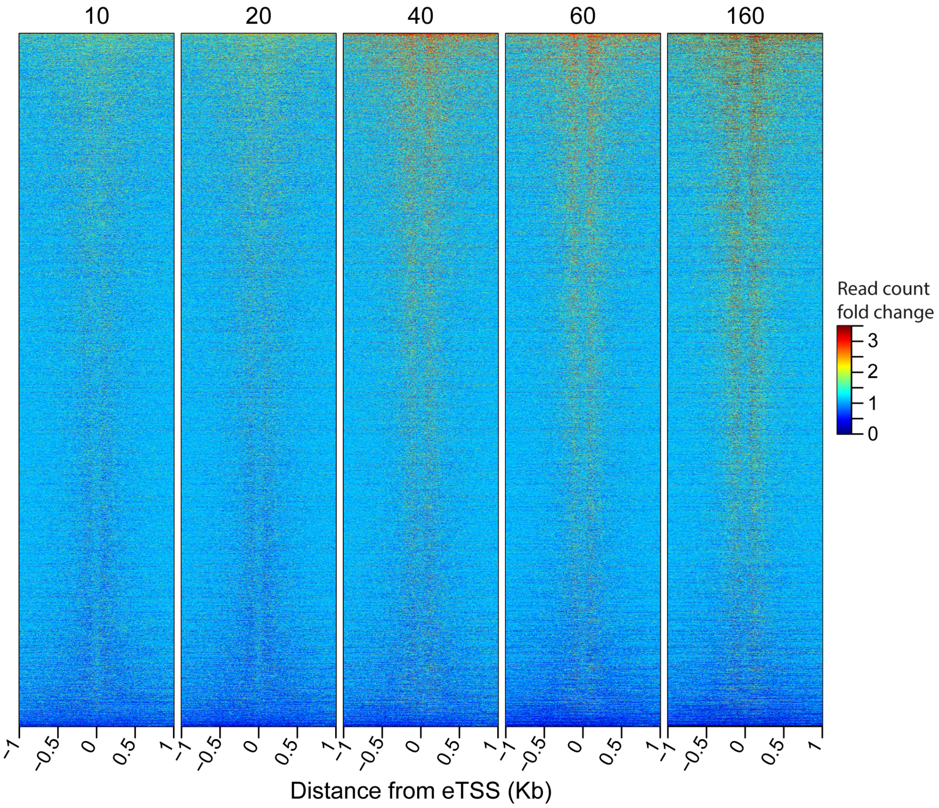


Figure 1.26: **Increased pausing at enhancers is pervasive.** Heatmap of fold-change in read counts relative to the 0-minute time point near all enhancers. Numbers at top indicate minutes after celastrol treatment.

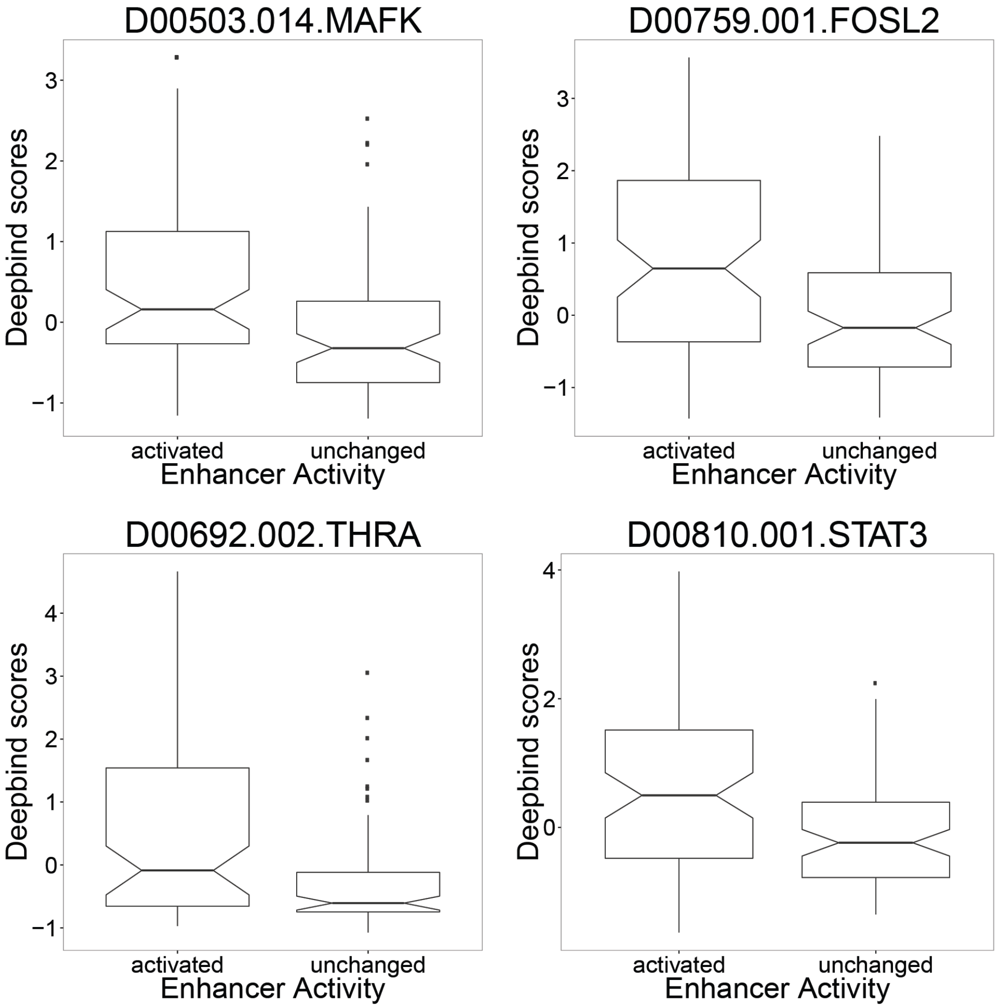


Figure 1.27: **A subset of transcription factors is associated with enhancer activation.** Activated enhancers are those that were not transcribed pre-celastrol treatment but were strongly up-regulated at some point in the time course whereas unchanged enhancers showed no change in expression throughout the whole time course (matched for average expression level). DeepBind scores are computed per TF per enhancer. The notch corresponds to median +/− $1.58 \cdot IQR/\sqrt{n}$, roughly a 95% confidence interval of the median.

1.7 Supplemental Tables

Table 1.1: **Distribution of marginal F-statistics and p-values for each motif in the TF regression using DeepBind scores.** F-statistics were computed per motif scores (there may be multiple motifs for the same TF), summing over the effects across all time points.

ID	TF	F	p
D00504.005	MAX	50.8	0.019
D00328.018	CTCF	30.3	0.032
D00471.002	HSF2	25.5	0.038
D00470.005	HSF1	23.7	0.041
D00619.003	RFX5	18.0	0.054
D00363.003	ELK4	15.6	0.062
D00760.003	FOS	13.9	0.069
D00660.005	SRF	13.3	0.072
D00587.002	POU2F1	12.5	0.076
D00789.003	NFYB	12.3	0.077
D00382.003	ETV4	12.0	0.079
D00347.005	E2F4	11.0	0.086
D00687.001	TFEB	10.8	0.087
D00384.002	ETV6	10.8	0.088
D00318.001	CEBPD	9.4	0.100
D00785.001	MYC	9.3	0.100
D00710.007	YY1	9.2	0.101
D00363.004	ELK4	8.6	0.109
D00535.004	NFE2	7.9	0.117
D00650.005	SP1	7.9	0.118
D00776.005	JUND	7.7	0.119
D00423.005	GMEB2	7.5	0.122
D00317.003	CEBPB	7.2	0.128
D00817.001	TBP	7.0	0.131
D00799.001	REST	7.0	0.132
D00379.001	ETV1	6.8	0.134
D00756.007	EZH2	6.4	0.142
D00560.003	NRL	6.4	0.143
D00821.001	TRIM28	6.2	0.146
D00559.001	NRF1	6.2	0.147
D00777.002	JUN	6.0	0.151
D00409.004	GABPA	6.0	0.152
D00700.001	USF1	5.9	0.154
D00616.002	RFX3	5.6	0.159
D00686.002	TFE3	5.5	0.162
D00822.001	UBTF	5.5	0.162
D00356.010	ELF1	5.5	0.164
D00504.002	MAX	5.3	0.168

Continued on next column

Continued on next column

Table 1.1 (Continued)

ID	TF	F	p
D00487.003	JDP2	5.3	0.169
D00808.004	SMC3	5.2	0.172
D00739.001	ATF3	5.1	0.175
D00796.001	RAD21	5.0	0.177
D00815.001	TAL1	4.9	0.179
D00813.009	TAF1	4.7	0.187
D00442.003	HMX3	4.7	0.189
D00501.003	MAFF	4.6	0.192
D00825.001	ZBTB33	4.5	0.193
D00501.004	MAFF	4.5	0.195
D00507.002	MEF2D	4.4	0.198
D00709.002	XBP1	4.4	0.199
D00768.001	GTF2F1	4.3	0.201
D00766.002	GATA2	4.3	0.203
D00816.001	TBL1XR1	4.3	0.204
D00672.001	TCF3	4.1	0.209
D00417.005	GCM1	4.1	0.211
D00748.001	CBX3	4.0	0.213
D00755.003	EP300	3.8	0.224
D00606.002	PRRX2	3.8	0.225
D00761.001	FOXA1	3.8	0.226
D00755.005	EP300	3.7	0.228
D00805.001	SIRT6	3.7	0.232
D00613.002	RAX	3.5	0.241
D00780.001	KDM5B	3.4	0.247
D00394.003	FOXD2	3.4	0.249
D00505.003	MEF2A	3.4	0.250
D00361.001	ELK3	3.3	0.252
D00433.003	HEY2	3.3	0.257
D00753.001	CTCFL	3.2	0.257
D00807.001	SMARCB1	3.2	0.260
D00292.001	ALX4	3.1	0.266
D00582.001	PITX3	3.1	0.269
D00700.006	USF1	3.0	0.271
D00765.001	GATA1	3.0	0.271
D00483.001	IRX5	3.0	0.274
D00537.001	NFIB	2.9	0.278
D00758.001	FOSL1	2.9	0.279
D00685.003	TFCP2	2.9	0.282
D00475.006	IRF3	2.8	0.287

Continued on next column

Table 1.1 (Continued)

ID	TF	F	p
D00503.014	MAFK	2.8	0.289
D00323.002	CPEB1	2.8	0.290
D00492.003	LEF1	2.8	0.292
D00317.009	CEBPB	2.7	0.299
D00714.003	ZBTB7A	2.6	0.304
D00353.005	EGR2	2.6	0.306
D00383.002	ETV5	2.5	0.312
D00546.003	NKX3-1	2.5	0.320
D00673.001	TCF4	2.5	0.321
D00395.002	FOXD3	2.4	0.322
D00824.001	WRNIP1	2.4	0.324
D00632.001	SHOX2	2.4	0.329
D00630.003	SCRT1	2.3	0.338
D00626.005	RXRA	2.3	0.342
D00356.005	ELF1	2.3	0.343
D00404.002	FOXO1	2.3	0.343
D00762.001	FOXA2	2.2	0.345
D00592.003	POU3F2	2.2	0.350
D00691.001	TGIF2	2.2	0.353
D00421.003	GLIS2	2.2	0.353
D00787.002	NFIC	2.1	0.358
D00581.002	PITX1	2.1	0.365
D00490.003	KLF16	2.0	0.371
D00468.003	HOXD8	2.0	0.373
D00660.007	SRF	2.0	0.373
D00779.001	KDM5A	2.0	0.376
D00538.001	NFIL3	2.0	0.381
D00678.001	TEAD3	1.9	0.391
D00559.006	NRF1	1.9	0.394
D00677.003	TEAD1	1.8	0.399
D00774.001	IKZF1	1.8	0.403
D00324.003	CREB3L1	1.8	0.404
D00818.003	TCF12	1.8	0.404
D00626.009	RXRA	1.8	0.411
D00409.003	GABPA	1.7	0.417
D00512.001	MEIS3	1.7	0.418
D00326.002	CREB3	1.7	0.422
D00601.003	PRDM4	1.7	0.424
D00398.001	FOXJ2	1.7	0.427
D00663.002	TBX15	1.7	0.430

Continued on next column

Continued on next column

Table 1.1 (Continued)

ID	TF	F	p
D00293.003	ARNTL	1.6	0.440
D00654.003	SPDEF	1.6	0.440
D00408.003	FOXP3	1.6	0.440
D00680.001	TEF	1.6	0.441
D00600.004	PRDM1	1.6	0.448
D00752.001	CTBP2	1.6	0.449
D00812.001	SUZ12	1.5	0.451
D00345.002	E2F2	1.5	0.452
D00328.003	CTCF	1.5	0.469
D00577.002	PAX9	1.5	0.471
D00523.003	MSC	1.4	0.472
D00782.001	MEF2C	1.4	0.473
D00529.004	MYBL2	1.4	0.480
D00529.001	MYBL2	1.4	0.481
D00405.003	FOXO3	1.4	0.483
D00653.003	SP8	1.4	0.487
D00488.003	KLF13	1.4	0.487
D00407.003	FOXO6	1.3	0.493
D00652.004	SP4	1.3	0.495
D00769.001	HDAC1	1.3	0.497
D00368.007	EN2	1.3	0.498
D00304.003	BATF3	1.3	0.499
D00763.001	FOXM1	1.3	0.499
D00321.003	CENPB	1.3	0.500
D00320.001	CEBPG	1.3	0.509
D00364.002	EMX1	1.3	0.513
D00547.003	NKX3-2	1.3	0.515
D00540.002	NFKB1	1.3	0.517
D00406.002	FOXO4	1.3	0.517
D00614.001	RFX2	1.2	0.520
D00552.006	NR2C2	1.2	0.526
D00631.002	SCRT2	1.2	0.532
D00750.001	CHD1	1.2	0.536
D00539.001	NFIX	1.2	0.536
D00788.001	NFYA	1.2	0.537
D00770.002	HDAC2	1.2	0.538
D00704.001	VDR	1.2	0.540
D00344.005	E2F1	1.1	0.548
D00508.004	MEIS1	1.1	0.552
D00430.003	HES7	1.1	0.557

Continued on next column

Continued on next column

Table 1.1 (Continued)

ID	TF	F	p
D00810.001	STAT3	1.1	0.559
D00744.001	BCLAF1	1.1	0.561
D00349.002	E2F8	1.1	0.570
D00652.003	SP4	1.1	0.570
D00575.003	PAX6	1.1	0.572
D00459.003	HOXC10	1.0	0.579
D00461.003	HOXC12	1.0	0.581
D00619.007	RFX5	1.0	0.586
D00783.001	MTA3	1.0	0.590
D00809.002	SP2	1.0	0.593
D00340.003	DMBX1	1.0	0.594
D00651.003	SP3	1.0	0.594
D00746.004	BHLHE40	1.0	0.598
D00429.003	HES5	1.0	0.602
D00335.003	DLX2	0.9	0.608
D00692.002	THRA	0.9	0.611
D00329.002	CUX1	0.9	0.613
D00689.003	TGIF1	0.9	0.616
D00416.005	GBX2	0.9	0.618
D00655.002	SPI1	0.9	0.625
D00381.003	ETV3	0.9	0.631
D00771.001	HDAC6	0.9	0.633
D00396.003	FOXG1	0.9	0.635
D00819.002	TCF7L2	0.9	0.640
D00435.003	HIC2	0.8	0.647
D00503.004	MAFK	0.8	0.648
D00348.003	E2F7	0.8	0.649
D00533.003	NFAT5	0.8	0.651
D00558.002	NR4A2	0.8	0.652
D00655.006	SPI1	0.8	0.664
D00370.003	ERF	0.8	0.671
D00522.002	MNX1	0.8	0.671
D00591.002	POU3F1	0.8	0.674
D00351.001	EGR1	0.8	0.676
D00299.003	ATF7	0.8	0.679
D00510.003	MEIS2	0.8	0.680
D00801.001	SAP30	0.7	0.693
D00794.047	POLR2A	0.7	0.693
D00542.005	NHLH1	0.7	0.701
D00366.003	EN1	0.7	0.703

Continued on next column

Table 1.1 (Continued)

ID	TF	F	p
D00424.002	GRHL1	0.7	0.703
D00611.001	RARG	0.7	0.707
D00710.002	YY1	0.7	0.715
D00740.002	BACH1	0.7	0.720
D00695.002	TP63	0.6	0.735
D00441.003	HMX2	0.6	0.739
D00759.001	FOSL2	0.6	0.748
D00412.002	GATA5	0.6	0.749
D00681.002	TFAP2A	0.6	0.752
D00775.001	JUNB	0.6	0.758
D00600.001	PRDM1	0.6	0.767
D00741.001	BATF	0.6	0.772
D00642.003	SOX18	0.5	0.778
D00541.001	NFKB2	0.5	0.779
D00565.002	ONECUT2	0.5	0.782
D00524.002	MSX1	0.5	0.786
D00754.003	E2F6	0.5	0.793
D00747.001	BRCA1	0.5	0.796
D00337.001	DLX4	0.5	0.803
D00351.006	EGR1	0.5	0.804
D00318.004	CEBPD	0.5	0.809
D00635.001	SMAD3	0.5	0.813
D00489.002	KLF14	0.5	0.814
D00519.002	MLXIPL	0.5	0.819
D00401.003	FOXK1	0.5	0.820
D00484.003	ISL2	0.5	0.826
D00516.002	MESP1	0.5	0.827
D00664.002	TBX19	0.4	0.829
D00705.003	VENTX	0.4	0.833
D00772.001	HMGN3	0.4	0.847
D00713.003	ZBTB49	0.4	0.850
D00583.002	PKNOX1	0.4	0.851
D00743.001	BCL3	0.4	0.852
D00662.003	TBR1	0.4	0.853
D00506.003	MEF2B	0.4	0.857
D00393.003	FOXC2	0.4	0.881
D00784.004	MXI1	0.3	0.885
D00595.001	POU4F1	0.3	0.888
D00495.003	LHX6	0.3	0.890
D00749.001	CCNT2	0.3	0.893

Continued on next column

Continued on next column

Table 1.1 (Continued)

ID	TF	F	p
D00792.001	PHF8	0.3	0.900
D00665.002	TBX1	0.3	0.902
D00599.006	POU6F2	0.3	0.902
D00767.001	GTF2B	0.3	0.903
D00802.001	SETDB1	0.3	0.905
D00681.004	TFAP2A	0.3	0.916
D00478.003	IRF7	0.3	0.923
D00344.002	E2F1	0.3	0.927
D00790.001	NR2F2	0.3	0.928
D00556.003	NR3C1	0.3	0.930
D00535.003	NFE2	0.3	0.937
D00432.003	HEY1	0.2	0.940
D00410.009	GATA3	0.2	0.950
D00351.009	EGR1	0.2	0.954
D00623.003	RUNX2	0.2	0.963
D00347.003	E2F4	0.2	0.965
D00552.002	NR2C2	0.2	0.969
D00517.002	MGA	0.2	0.970
D00346.003	E2F3	0.2	0.971
D00410.003	GATA3	0.2	0.975
D00305.003	BCL6B	0.2	0.976
D00418.003	GCM2	0.2	0.979
D00650.007	SP1	0.2	0.979
D00745.001	BDP1	0.1	0.987
D00791.001	PBX3	0.1	0.988
D00505.006	MEF2A	0.1	0.990
D00797.001	RBBP5	0.1	0.993
D00536.003	NFIA	0.1	0.994
D00793.002	PML	0.1	0.997
D00475.003	IRF3	0.1	0.997
D00528.001	MYBL1	0.1	0.997
D00679.004	TEAD4	0.1	0.998
D00431.002	HESX1	0.0	1.000

Continued on next column

Concluded

Table 1.2: **Distribution of the marginal F-statistics and p-values for each motif in the TF regression using ChIP-seq scores.** F-statistics were computed per motif scores (there may be multiple motifs for the same TF), summing over the effects across all time points.

TF	F	p
POLR2A	227.6	0.004
MYC	77.6	0.013
RFX1	67.4	0.015
TBP	49.3	0.020
CCNT2	45.8	0.022
UBTF	43.6	0.023
RAD21	40.8	0.024
E2F4	35.1	0.028
PHF8	34.8	0.028
SP1	32.5	0.030
PML	30.3	0.032
MXI1	30.2	0.032
SUZ12	25.3	0.039
CHAMP1	25.2	0.039
RCOR1	24.2	0.040
MLLT1	23.5	0.041
MAX	23.0	0.042
SAP30	21.9	0.044
C11orf30	21.0	0.046
NFYA	20.5	0.047
HDAC1	18.5	0.052
ZBTB40	16.6	0.058
ZBTB33	16.1	0.060
ESRRA	15.9	0.060
NFYB	15.7	0.061
ZNF318	15.5	0.062
GTF2B	12.9	0.074
ELF1	12.9	0.074
ATF1	12.8	0.074
RAD51	12.3	0.077
TRIM24	12.0	0.079
YY1	11.8	0.080
L3MBTL2	11.4	0.083
SMAD5	11.2	0.084
COPS2	10.6	0.089
RBBP5	10.5	0.090
CREM	9.7	0.097
IRF2	9.6	0.098

Continued on next column

Continued on next column

Table 1.2 (Continued)

TF	F	p
CEBPZ	9.2	0.101
TAF1	9.0	0.104
BMI1	8.1	0.114
MAZ	8.1	0.114
GABPA	7.6	0.122
NFE2	7.5	0.124
BACH1	7.2	0.127
TEAD4	7.1	0.130
YBX1	7.0	0.130
USF1	7.0	0.131
TARDBP	6.6	0.139
GATA1	6.6	0.139
NRF1	6.5	0.140
HDAC2	6.2	0.145
FOXK2	6.2	0.146
ATF3	6.2	0.146
CHD1	6.2	0.146
GTF3C2	6.1	0.149
PKNOX1	6.0	0.151
CTCFL	6.0	0.151
SIN3A	5.9	0.153
SMARCA4	5.8	0.156
SMARCA5	5.7	0.157
GATA2	5.7	0.158
SREBF1	5.6	0.159
MAFK	5.5	0.163
SIX5	5.5	0.163
TAF7	5.4	0.164
ZNF143	5.4	0.166
SP2	5.0	0.177
EP300	4.9	0.179
BDP1	4.9	0.181
CTCF	4.9	0.181
RNF2	4.8	0.184
SRF	4.8	0.185
KDM5B	4.7	0.186
ZBTB7A	4.7	0.188
ZHX1	4.5	0.196
FOS	4.1	0.210
TRIM28	4.0	0.216

Continued on next column

Table 1.2 (Continued)

TF	F	p
EGR1	4.0	0.216
ZBTB11	3.9	0.218
CEBPB	3.9	0.220
E2F6	3.8	0.223
NFRKB	3.8	0.226
CDC5L	3.7	0.228
DPF2	3.6	0.236
GTF2F1	3.5	0.242
SETDB1	3.4	0.246
EZH2	3.3	0.251
HDGF	3.1	0.268
SMARCB1	2.9	0.278
CBX3	2.9	0.284
RUNX1	2.9	0.284
POLR3G	2.8	0.289
HDAC6	2.7	0.302
ARID3A	2.6	0.303
NR2F6	2.6	0.309
NR2C2	2.6	0.309
CHD2	2.4	0.326

Continued on next column

TF	F	p
SPI1	2.4	0.328
NBN	2.4	0.332
DDX20	2.2	0.348
IKZF1	2.2	0.349
REST	2.2	0.356
MEF2A	2.1	0.366
STAT5A	2.0	0.374
TCF7	2.0	0.375
MAFF	2.0	0.379
BCLAF1	1.9	0.386
POLR3A	1.9	0.389
BHLHE40	1.9	0.391
MIER1	1.9	0.396
YBX3	1.8	0.407
KDM1A	1.8	0.410
MYNN	1.7	0.416
THAP1	1.6	0.449
MCM3	1.5	0.461
JUN	1.5	0.467
ZEB2	1.5	0.469

Continued on next column

Table 1.2 (Continued)

TF	F	p
SMC3	1.4	0.480
ETV6	1.4	0.487
KLF16	1.3	0.498
HES1	1.3	0.504
TAL1	1.3	0.505
BCOR	1.2	0.522
TBL1XR1	1.2	0.535
ELK1	1.2	0.536
SMARCE1	1.1	0.570
CBX5	1.0	0.577
NR2F2	1.0	0.585
ZMIZ1	0.9	0.609
FOSL1	0.9	0.618
MCM7	0.9	0.642
MBD2	0.8	0.645
ZNF24	0.8	0.648
DEAF1	0.8	0.650
MITF	0.8	0.668
MCM5	0.8	0.671
SMAD2	0.8	0.678
ARNT	0.7	0.692
BRF2	0.7	0.714
JUNB	0.7	0.717
NCOA1	0.6	0.735
ZNF274	0.6	0.742
KAT8	0.6	0.753
JUND	0.6	0.760
MTA2	0.4	0.878
SIRT6	0.4	0.880
USF2	0.3	0.885
ZNF263	0.2	0.960
LEF1	0.1	0.983
CBX1	0.1	0.987
ZKSCAN1	0.1	0.990
RFX5	0.0	1.000
HMBOX1	0.0	1.000
HMGN3	0.0	1.000
KDM4B	0.0	1.000

Continued on next column

Concluded

CHAPTER 2

IS A SUPER-ENHANCER GREATER THAN THE SUM OF ITS PARTS?

Note: With the exception of the preface, the introduction, and few minor changes, this chapter contains the same text as the previously published work "Is a super-enhancer greater than the sum of its parts?" in *Nature Genetics* (Volume 49, Number 1).

2.1 Preface

This chapter is somewhat unusual, both in its origin and its brevity. It was conceived of based on a comparison between two papers, Shin et al. (2016) and Hay et al. (2016), during a lab journal club. This discussion convinced Dr. Brad Gulko to do some preliminary analysis the following weekend, and seeing promising results, the three first authors jointly performed the work for the super-enhancer paper. In early 2018, based on requests from other groups for help performing similar analysis, the author of this thesis wrote the superEnhancerModelR R package (Dukler, 2018).

2.2 Introduction

The first enhancer was described in 1980, based on deletion studies of a 72-bp sequence in Simian Virus 40 genome (Benoist and Chambon, 1981; Gruss et al., 1981). Shortly thereafter, an example was found in mammalian cells and the ability of enhancers to act in a distance independent fashion was established (Banerji et al., 1983; Gillies et al., 1983; Mercola et al., 1983). Given that enhancers could be quite far from the gene that they regulated, locating them could be a major challenge. The first next major advance on this front came during the early short read sequencing era, when work from Gregory Crawford's and Bing Ren's groups discovered that the histone H3K4me1 modification and p300 binding were indicative of enhancer function (Heintzman et al., 2007). Shortly thereafter additional genomic signatures of enhancers were identified, which further facilitated genome-wide enhancer discovery (Heintzman et al., 2009; De Santa et al., 2010; Kim et al., 2010).

The term "super-enhancer" was coined by Whyte et al. (2013) to describe their observation that a small fraction of enhancers were strongly enriched for binding master regulators (Oct4, Sox2, Nanog) and Mediator (Med1), and that these elements seemed to regulate cellular identity. Additional work characterizing super-enhancers suggested that they also played an important role in disease (Hnisz et al., 2013, 2015; Vahedi et al., 2015). As papers about super-enhancers proliferated (from 15 in 2013 to 182 in 2016[1]), so did ways of detecting and defining them. For example, Hnisz et al. (2013) used H3K27Ac signal, not TF binding, to locate individual enhancers, and H3K27Ac enrichment instead of Med1 enrichment to filter for super-enhancers. Independently Parker et al. (2013) described a similar concept, the "stretch-enhancer", characterized by long regions of chromatin (≥ 3kb) with an enhancer signature. Like super-enhancers, "stretch-enhancers" showed a tendency to be cell type specific and enrich for disease variants (Parker et al., 2013; Quang et al., 2015). Despite their similarities "super-enhancers" and "stretch-enhancers" had several notable differences including there being an order of magnitude more "stretch-enhancers" than "super-enhancers", leading to debate in the community about whether ei-

[1] Based on Web of Science search for term "super-enhancer"

ther concept represented a truly novel mechanism of gene regulation beyond the existing enhancer literature (Pott and Lieb, 2015).

By this point a large body of evidence had accumulated that enhancers played a key role in both development and disease (Smith and Shilatifard, 2014). Combined with recent developments in gene editing technology (TALENs, CRISPR), this led to a wave of studies attempting to characterize these elements *in vivo* (Canver et al., 2015; Shin et al., 2016; Hay et al., 2016; Moorthy et al., 2017). Two of these studies, Shin et al. (2016) and Hay et al. (2016), focused on the *Wap* and α-globin respectively, which were well characterized and thus valuable model systems. The α-globin locus contains HBZ (ζ-hemoglobin), HBA2 (α2-hemoglobin), HBA1 (α1-hemoglobin) and HBQ1 (θ-hemoglobin) genes which are specifically expressed in erythroid cells. Previous work on this region had shown evidence of multiple distal cis-regulatory elements under selective pressure (Hughes et al., 2005) that looped to α-globin promoters (Hughes et al., 2014). The *Wap* super-enhancer is able to induce the *Wap* gene, which code for a core whey protein (Simpson et al., 2000), over 1000-fold specifically in mammary tissue during pregnancy (Burdon et al., 1991). Given the controversy over the nature of super-enhancers, the enhancer-dependant inducibility and cell type specificity of both loci made them obvious candidates for dissecting the "super-enhancer" concept.

2.3 Letter to the editor

The recent back-to-back articles by Hay et al. (2016) and Shin et al. (2016) both addressed the important question of how the constituent enhancers of a so-called "super-enhancer" combine to activate the expression of a target gene. Super-enhancers are collections of closely spaced genomic regions that exhibit hallmarks of enhancers, such as binding by the Mediator complex and acetylation of histone H3 at lysine 27 (H3K27ac)(Hnisz et al., 2013; Whyte et al., 2013; Heinz et al., 2015). As these authors noted, there is continuing controversy over whether super-enhancers genuinely represent a new paradigm in transcriptional regulation or whether they may essentially just be clusters of conventional enhancers that together produce a strong transcriptional response(Pott and Lieb, 2015).

At the heart of this question is whether the activity of a super-enhancer is simply given by the sum of its constituent enhancers—that is, whether it is *additive*—or whether these components instead exhibit some kind of synergy. Indeed, this question of additivity is of general interest, whether or not super-enhancers are qualitatively distinct from other loci. Hay et al. (2016) and Shin et al. (2016) addressed this question by carefully dissecting the highly expressed α-globin and *Wap* loci , respectively, andmeasuring the reductions in gene expression resulting from several individual and combined knockouts of constituent enhancers. Both articles described highly variable effects on gene expression from different individual knockout experiments, and both reported that it was necessary to disable multiple enhancers to abolish, or nearly abolish, expression. On the question of additivity, however, the two articles reached strikingly different conclusions: Hay et al. reported that the constituent enhancers at the α-globin locus acted "independently and in an additive fashion," whereas Shin et al. reported that their observations of the *Wap* super-enhancer supported a "temporal and functional hierarchy" of constituent enhancers that is presumably non-additive.

It was notable that neither of these articles offered a precise definition for “additivity” or “hierarchy”. Moreover, neither article explicitly compared a null hypothesis of additivity against an alternative hypothesis. In reviewing these two works, we became interested in the various ways in which a super-enhancer’s activity could plausibly be modeled using a linear function of the activity of its constituent enhancers, possibly together with a simple nonlinear “link” function(Nelder and Wedderburn, 1972), and in whether the data would allow a null hypothesis of such generalized linearity to be formally rejected. Here we show, by reanalyzing these two data sets, that they are both consistent with a generalized linear model that has a simple biophysical interpretation and does not require any hierarchy or synergy among constituent enhancers. Thus, we argue that it still remains to be demonstrated that a super-enhancer is greater than the sum of its parts.

Perhaps the simplest linear model would assume each constituent enhancer makes an additive contribution directly to the expression level of the target gene, such as might be the case if the constituent enhancers separately contribute to transcription. (This appears to be the model that Hay et al. (2016) had in mind.) Specifically, let us define the “activity” of the super-enhancer by the affine (linear plus constant) function,

$$A(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n, \tag{2.1}$$

where $\mathbf{x} = (x_1, \ldots, x_n)'$ is a vector of binary variables indicating whether each constituent enhancer x_i is present ($x_i = 1$) or absent ($x_i = 0$) and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$ is a corresponding vector of real-valued coefficients, with β_0 as an intercept term. Then, we can model measurements of the expression of the target gene, $R(\mathbf{x})$, as a combination of this activity function and a variable ϵ representing some combination of biological and experimental noise.

Thus, we write,

$$R(\mathbf{x}) = A(\mathbf{x}) + \epsilon. \qquad \textbf{(additive model)} \qquad (2.2)$$

In practice, we consider alternative noise models and find that a log-normal model fits the data best for all of the models that we consider (see Supplementary Note for details).

Another plausible scenario is that the constituent enhancers combine multiplicatively, rather than additively, in determining $R(\mathbf{x})$. This multiplicative relationship might be expected, say, if the constituent enhancers act to promote transcription in a sequential manner, with each step having the opportunity to amplify or dampen the outputs of previous steps. This relationship can be captured simply by making $R(\mathbf{x})$ an exponential, rather than an additive, function of the activity $A(\mathbf{x})$. Because the scale of $A(\mathbf{x})$ is determined by free parameters, the base associated with the exponent is unimportant. By convention, we use base e and write,

$$R(\mathbf{x}) = e^{A(\mathbf{x})} + \epsilon. \qquad \textbf{(linear-exponential model)} \qquad (2.3)$$

Equation 2.3 can be considered a generalized linear model with inverse link function e^x (in the language of GLMs (Nelder and Wedderburn, 1972)). Importantly, it is fully determined by a linear activity function, with no explicit consideration of interactions between the constituent enhancers.

Notably, this model can be given an alternative biophysical interpretation.

Let us assume a physical system with two broadly defined "states," a low-energy state associated with active transcription and a higher-energy baseline state. (The model is abstract: in reality, these "states" may each correspond to large ensembles of particular configurations of molecules.) Furthermore, let us interpret $A(\mathbf{x})$ as a measure of the reduction in energy of the transcription-associated state relative to the baseline state. Statistical mechanics tells us that the occupancy of the low-energy state should be given by a Boltzmann distribution and be proportional to $e^{A(\mathbf{x})}/Z$, where Z is the partition function. If we further assume that the system is far from its optimum, then the occupancy of the low-energy state will be approximately proportional to $e^{A(\mathbf{x})}$. Equation 2.3 can therefore be interpreted as the model that results from assuming transcription is proportional to occupancy of the low-energy state in this suboptimal regime.

This physical interpretation, with a two-state system and a linear energy function, leads naturally to a third generalized linear model. In this case, we abandon the "suboptimal" approximation and consider the full Boltzmann distribution for the system(Lässig, 2007; Phillips, 2015). In the two-state model, we can explicitly calculate the partition function Z and write, $e^{A(\mathbf{x})}/Z = e^{A(\mathbf{x})}/\left(1 + e^{A(\mathbf{x})}\right) = 1/\left(1 + e^{-A(\mathbf{x})}\right)$, which is known as a logistic function of $A(\mathbf{x})$. Thus, we can fully describe the fraction of time the low-energy state is occupied using a generalized linear model with the logistic function as the inverse link function. Assuming again that gene expression is proportional to the occupancy of the low-energy state, we write,

$$R(\mathbf{x}) = \frac{\gamma}{1 + e^{-A(\mathbf{x})}} + \epsilon, \qquad \textbf{(linear-logistic model)} \qquad (2.4)$$

where γ defines the maximum expected level of gene expression (for a similar model applied to enhancers, see Crocker et al. (2016)). Equation 2.4 will behave similarly to equation 2.3 when $A(\mathbf{x})$ is far from its optimum but it will capture the phenomenon of diminishing returns in transcriptional output as the energetics of productive transcriptional elongation approach an optimum and gene expression is limited by other features of the system (saturation).

We fitted these three models (equations 2.2–2.4) to the raw data from Hay et al. (2016) and Shin et al. (2016) by maximum likelihood using a numerical algorithm for optimization. The data consisted of all replicates for each tested configuration (wild type and knockout) of the three constituent enhancers of the *Wap* super-enhancer and the five constituent enhancers of the α-globin super-enhancer (see Supplementary Note for complete details). We compared the goodness-of-fit of the models using the Bayesian Information Criterion (BIC), which penalizes more complex models for their additional parameters. (Here, the linear-logistic model has one additional parameter, γ.)

For the α-globin data set (Hay et al., 2016), for which the authors claimed additivity, we found that the additive model did indeed fit the data fairly well (Figure 2.1A). Nevertheless, the linear-logistic model was preferred over the additive model according to the BIC, despite its additional parameter. For the *Wap* data set (Shin et al., 2016), the linear-logistic model is the best-fitting model by a substantial margin. Thus, for both of these data sets, the linear-logistic model explains the observed data better than any other generalized linear model (Figure 2.1B&C), and therefore is a better null model than the additive model. Notably, the linear-logistic model explains both data sets well despite several important differences between the two loci (e.g., the *Wap* component enhancers are substantially more tightly clustered and closer to the TSS than those for α-

globin) and between the knock-out strategies used (Shin et al. deleted STAT5-binding sites whereas Hay et al. deleted larger DNase-I hypersensitive regions), which underscores the flexibility and generality of this simple model.

But do the data of Shin et al. (2016) for the *Wap* super-enhancer truly support something more complex than a generalized linear relationship, as the authors seem to claim? We attempted to address this question quantitatively in our framework by introducing interaction terms for the two pairs of constituent enhancers that were simultaneously knocked out in that study (ΔE1a/ΔE2 & ΔE2/ΔE3). We found that models allowing for interactions between constituent enhancers do have slightly higher likelihoods than the simple linear-logistic model, as they must, but, according to the BIC, these improvements are not sufficient to justify the use of an additional parameter (Figure 2.1D; see Supplementary Note for details). Thus, we find not only that the linear-logistic model fits both the α-globin and *Wap* data sets reasonably well, but also that this model cannot confidently be rejected in favor of one that allows for interactions between constituent enhancers.

It is possible, of course, that interactions between component enhancers do occur in reality, but the data collected so far are insufficiently abundant or precise to reject a generalized linear null model. In addition, our models are limited in that they address only the knockout data from these studies. In particular, our models do not address Shin et al.'s observation that the E1 enhancer is occupied by key transcription factors first during pregancy, suggesting possible non-additivity in temporal establishment of the *Wap* super-enhancer, if not in its subsequent regulatory behavior. Finally, it is worth emphasizing that our abstract modeling approach provides no direct mechanistic insights into tran-

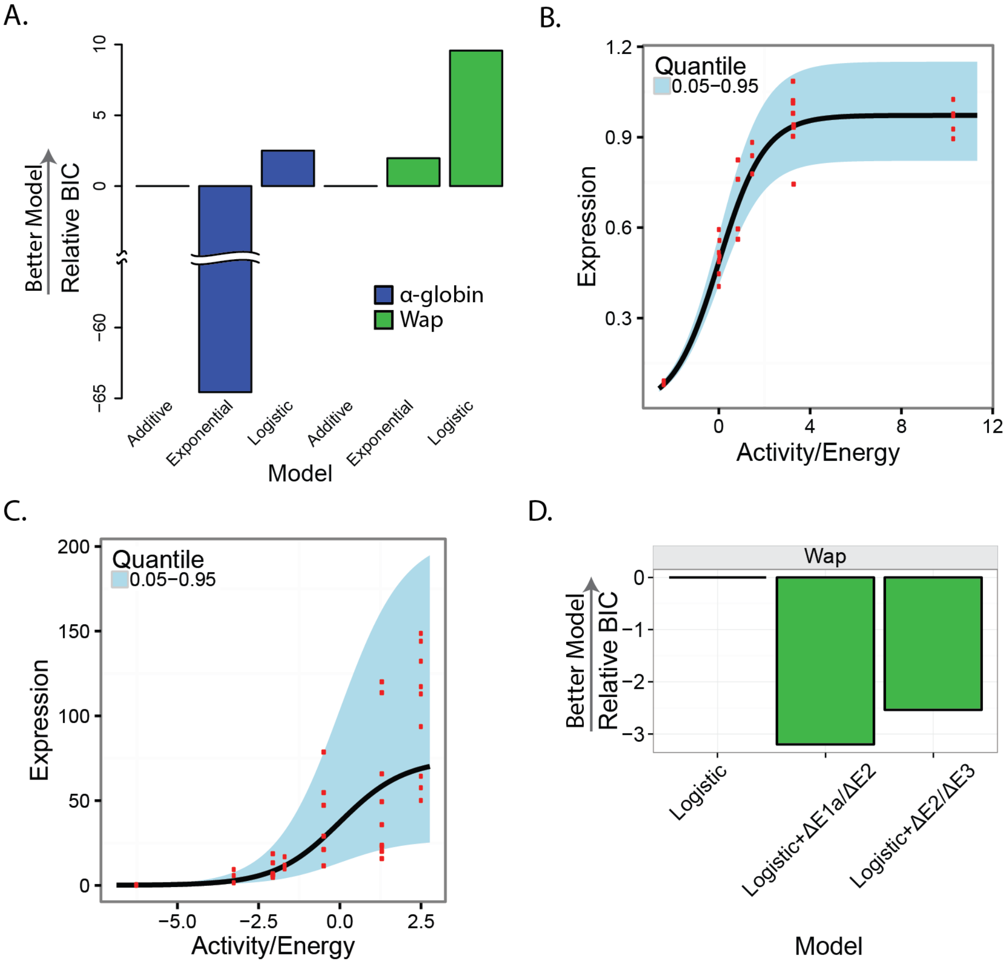


Figure 2.1: **Comparing model fits for the α-globin and *Wap* super-enhancers.**(A) Model fit for the α-globin (blue) and *Wap* (green) data sets, measured as the Bayesian Information Criterion (BIC) for the additive model minus the BIC for the additive (0 by definition), linear-exponential, and linear-logistic models. (B) Predicted expression at the α-globin locus (black line, with blue boundaries indicating 0.05–0.95 quantiles) under the best-fitting linear-logistic model. The actual data points are shown in red, with replicates aligned vertically. (C) Same for the *Wap* locus. In (B) and (C), the wild-type configuration appears as the collection of data points farthest to the right. Notice that the definition of the x-axis depends on the linear coefficients estimated for each model. (D) Model fit, measured by BIC, for the *Wap* data set for the linear-logistic model shown in (A) relative to alternative models with interaction terms for either E1a & E2 or E2 & E3

scriptional regulation at either of these loci. Nevertheless, we have shown that the observed knockout data for both of these super-enhancers can be explained fairly well by a very simple generalized linear model, and this observation can at least constrain the family of possible mechanistic models. More broadly, we argue that the transcription field would benefit from clearer definitions of null models and more rigorous criteria for rejecting them before concluding that complex behaviors occur.

Note: The computer code originally developed for this analysis is available on GitHub (`https://github.com/CshlSiepelLab/super-enhancer-code`). Additionally, since the publication of the original article an R package has been created and is available at `https://github.com/ndukler/superEnhancerModelR`.

2.4 Supplemental Materials

2.4.1 Model design

We designed three models (additive, linear-exponential, and linear-logistic) to predict gene expression, each as a transformation of an affine function $A(\mathbf{x})$. With x_i as indicator variables for the presence of individual enhancers and β_i as the coefficients that estimate the relative enhancer contributions to gene expression, the activity function $A(\mathbf{x})$ is written:

$$A(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n,$$

Each model predicts gene expression $T(\mathbf{x})$ as a monotonically increasing function of $A(\mathbf{x})$: $T(\mathbf{x}) = g^{-1}(A(\mathbf{x}))$, where $g()$ is a link function and $g^{-1}()$ is its inverse (following the conventions of generalized linear models(Nelder and Wedderburn, 1972). The measured expression level, $R(\mathbf{x})$, is then assumed to be a combination of the true level $T(\mathbf{x})$ and some noise ϵ (see next section). The additive model is simply defined by the identity inverse link function:

$$T(\mathbf{x}) = A(\mathbf{x}).$$

The linear-exponential model is given by:

$$T(\mathbf{x}) = e^{A(\mathbf{x})}.$$

Finally, the linear-logistic model is given by:

$$T(\mathbf{x}) = \frac{\gamma}{1 + e^{-A(\mathbf{x})}}.$$

2.4.2 Error models

For each generalized linear model we specified normal and log-normal error models. The normal error model is defined by assuming that:

$$R(\mathbf{x}) = T(\mathbf{x}) + \epsilon,$$

where the noise ϵ is normally distributed with mean zero and variance σ^2, that is, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This is equivalent to saying that a measurement of gene expression for a particular value of $\mathbf{x}$, $R(\mathbf{x})$, is assumed to be normally distributed with mean $T(\mathbf{x})$ and variance σ^2. Thus, the density function for $R(\mathbf{x})$ is:

$$f(R(\mathbf{x})) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(R(\mathbf{x}) - T(\mathbf{x}))^2}{2\sigma^2}\right].$$

By contrast, the log normal error model assumes that the noise is normally distributed on a log scale, that is,

$$\log R(\mathbf{x}) = \log T(\mathbf{x}) + \epsilon,$$

where, again, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This is equivalent to saying that $R(\mathbf{x})$ has a log-normal distribution with density function,

$$f(R(\mathbf{x})) = \frac{1}{R(\mathbf{x})\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log R(\mathbf{x}) - \log T(\mathbf{x}))^2}{2\sigma^2}\right].$$

We found that the log-normal error model fit the data substantially better than the normal error model in all cases (Supplemental Fig. **??**), and we used it for all results discussed in the main text. Note that measuring error on a log scale, as in the log-normal model, makes sense when PCR amplification makes a large contribution to measurement error, as may be the case for the data from Shin et al. (2016).

2.4.3 Interaction Parameters

Shin et al. (2016) proposed a "functional hierarchy within the STAT5-driven *Wap* super-enhancer". Within our framework we defined hierarchical relationship between two enhancers as indicator variables whose state depended jointly on the presence of both enhancers. To test Shin et al.'s claims, we defined a set of augmented models that included additional indicator variables x_{ij} (defined for $i < j$) where:

$$\begin{cases} x_{ij} = 1 & \text{if } x_i = 1 \wedge x_j = 1 \\ x_{ij} = 0 & \text{otherwise} \end{cases}$$

We created two linear-logistic models with log-normal error, one that included x_{12}, and one that included x_{23}. We did not fit a separate model for x_{13} since there was no data for an $E1 - E3$ deletion. Each of these models had five centrality (non-error) parameters. Given that the dataset for the *Wap* super-enhancer only contains six conditions, we could not fit multiple interaction terms at once without over-fitting.

2.4.4 Model optimization

For all three models (additive, linear-exponential, and linear-logistic) we fit two error models, normal and log-normal, by maximum likelihood. All models were fit in R using the differential evolutionary algorithm package DEoptim (Mullen et al., 2011). The variance parameter σ was estimated as well as the coefficients $(\beta_0, \ldots, \beta_n)$. All replicates were considered simultaneously. A population with 20 times as many members as there were parameters was created and run for 10,000 steps. The best solution was then further optimized by running gradient descent (L-BFGS-B) until convergence. The Bayesian Information Criterion (BIC) was computed for each model using the formula:

$$\text{BIC} = -2\log(L) + k\log(N)$$

where L is the likelihood of the data with the parameters obtained via optimization, k is the number of free parameters, and N is the number of data points.

2.5 Supplemental Figures

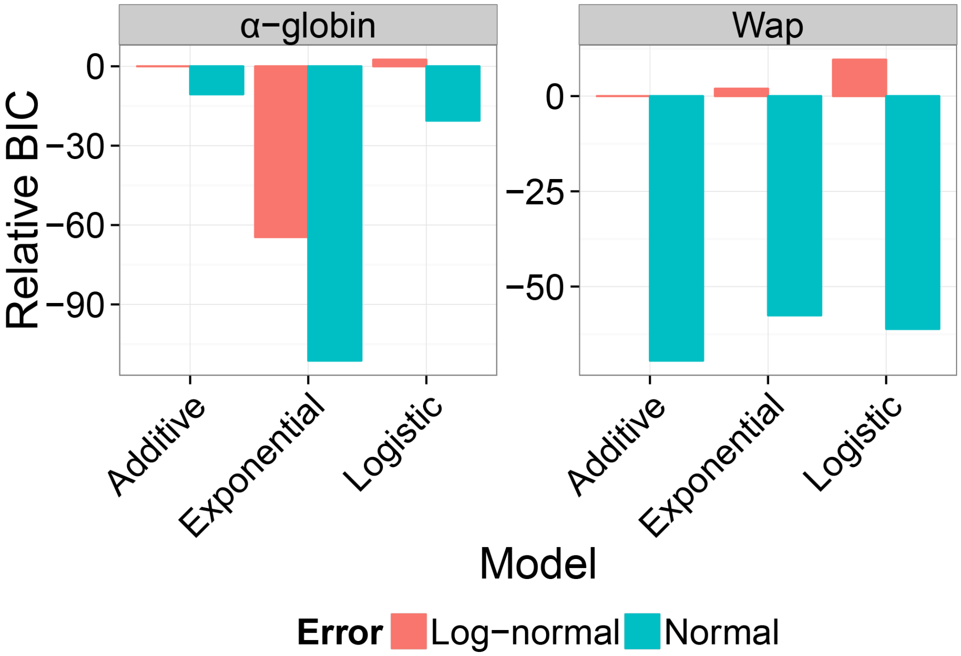


Figure 2.2: **Bayesian Information Criterion (BIC) for all models.** (A) α-globin locus. (B) *Wap* locus.

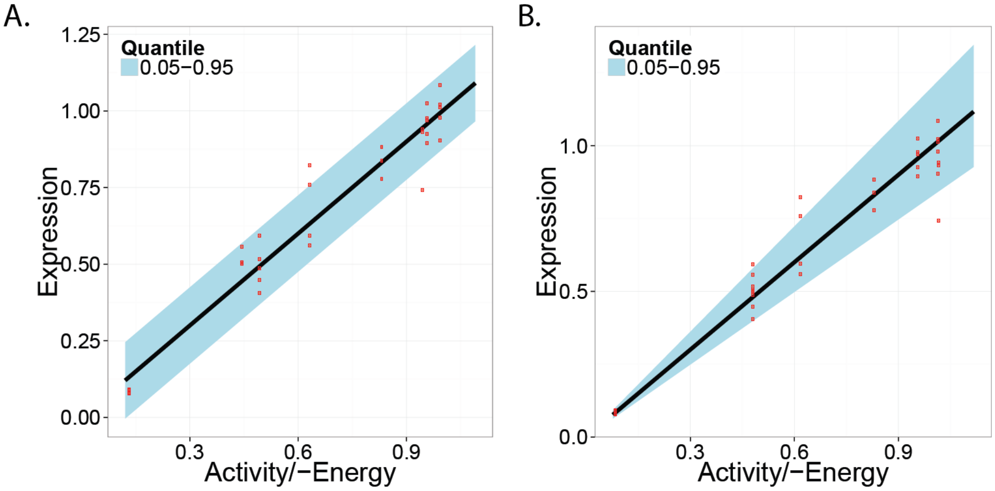


Figure 2.3: **Additive model with normal and log-normal error models fit to α-globin locus data.** (A) Normal error model. (B) Log-normal error model.

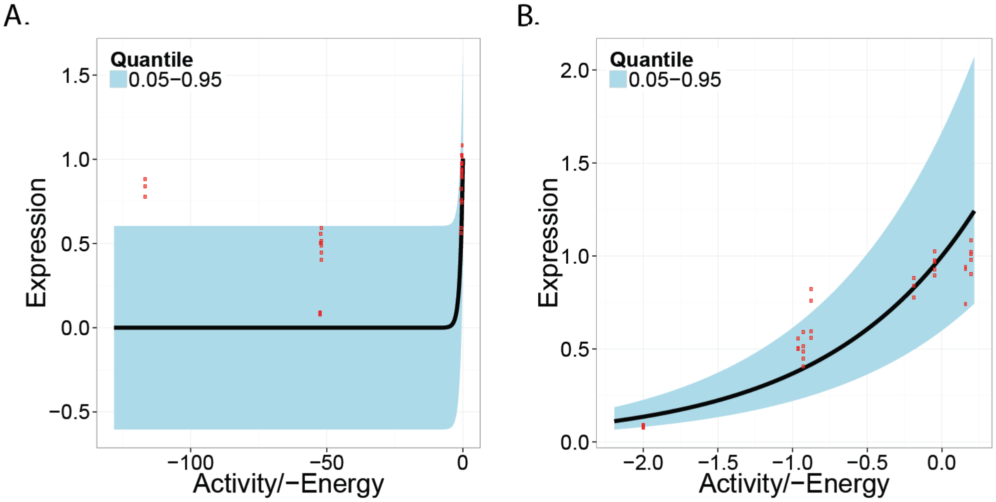


Figure 2.4: **Linear-exponential model with normal and log-normal error models fit to α-globin locus data.** (A) Normal error model. (B) Log-normal error model.

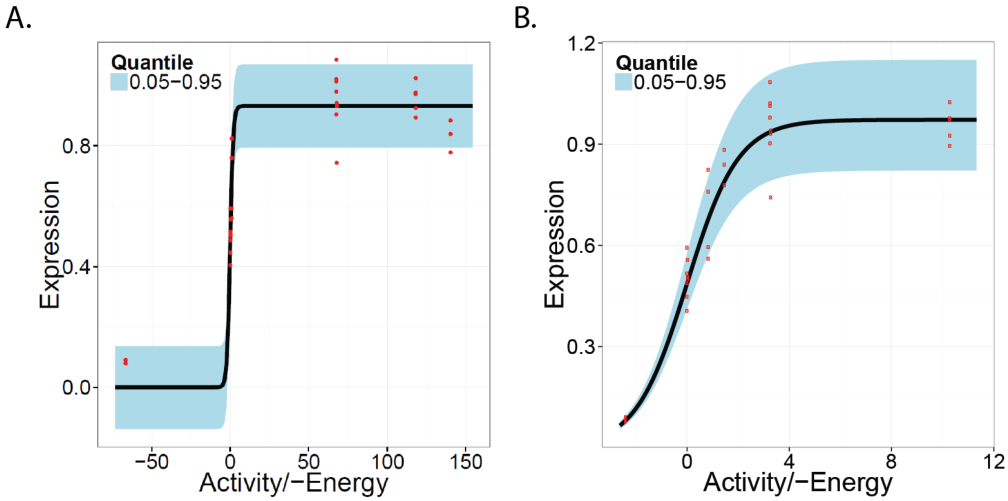


Figure 2.5: **Linear-logistic model with normal and log-normal error models fit to α-globin locus data.** (A) Normal error model. (B) Log-normal error model.

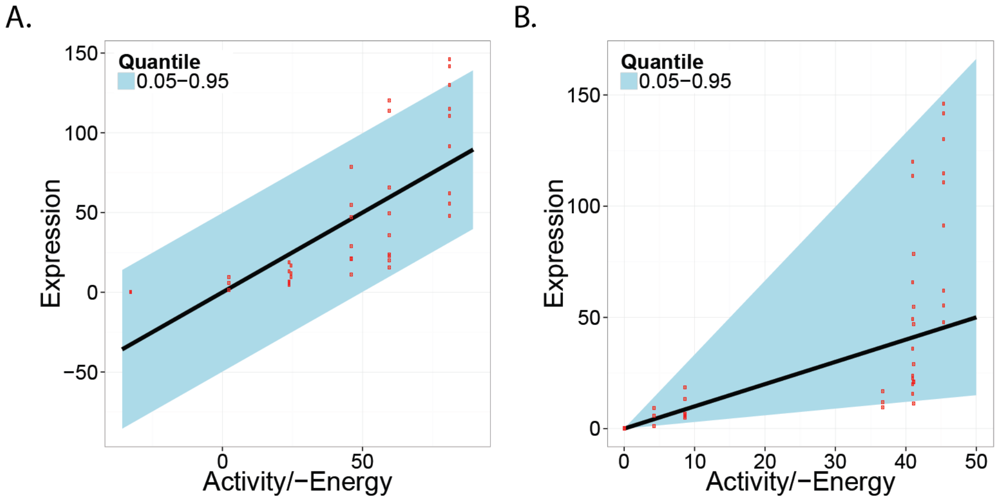


Figure 2.6: **Additive model with normal and log-normal error models fit to *Wap* locus data.** (A) Normal error model. (B) Log-normal error model.

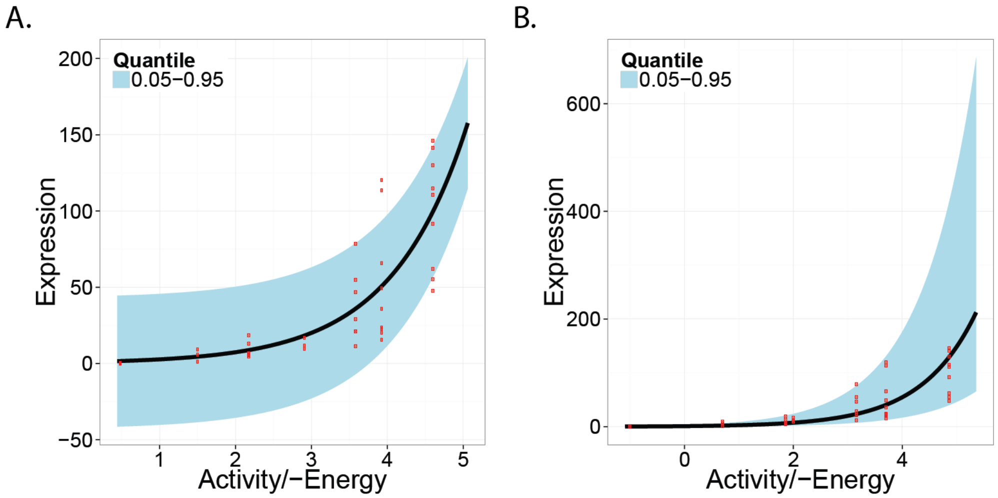


Figure 2.7: **Linear-exponential model with normal and log-normal error models fit to *Wap* locus data.** (A) Normal error model. (B) Log-normal error model.

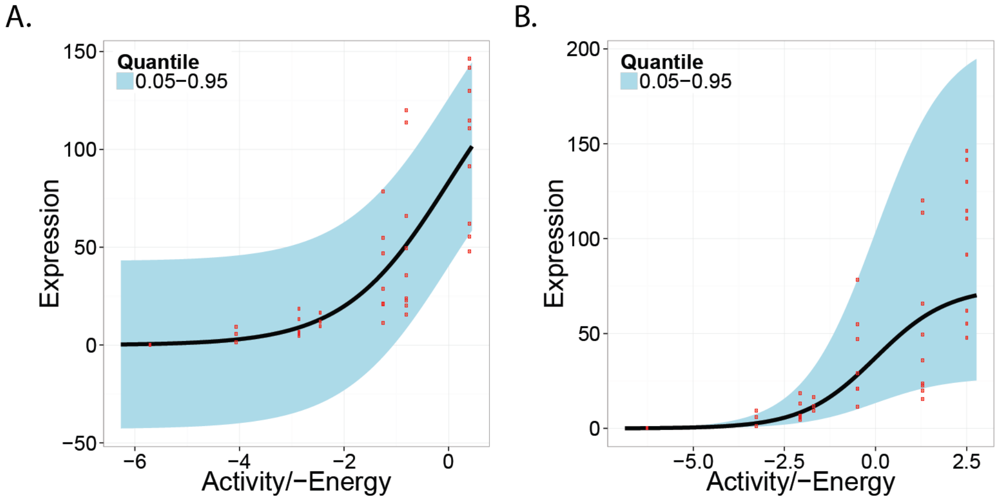


Figure 2.8: **Linear-logistic model with normal and log-normal error models fit to *Wap* locus data.** (A) Normal error model. (B) Log-normal error model.

CHAPTER 3

CHARACTERIZING THE GENOMIC DETERMINANTS OF CIS-REGULATORY ELEMENT EVOLUTION WITH PHYLOGENETIC MODELS

3.1 Introduction

Cis-regulatory elements (CREs) govern gene regulation and thereby phenotype to a large degree, accounting for the vast majority ($\backsim$ 80%) of narrow sense trait heritability (Gusev et al., 2014). As such, the evolutionary processes that govern the turnover of these elements have been of intense interest at both mechanistic and epigenetic levels.

Work using signatures of inter-species sequence conservation in mammals to locate CREs has estimated that at least 5% of the human genome has undergone purifying selection (Lindblad-Toh et al., 2011). However these approaches suffer from two main limitations: (1) they are unable to resolve the tissue(s) in which the CRE is active and (2) they have limited power to locate where the constrained regions actually are, especially in non-coding regions where selection may be weaker and act on shorter regions. Regions under positive selection have also been of great interest, with studies in the human lineage finding human-specific accelerated regions are mostly in non-coding regions (Pollard et al., 2006), and similar work in eutherian mammals finding evidence of acceleration in enhancers associated with neuronal development (Holloway et al., 2016).

However, it has been shown that comparative sequence based methods may be missing a majority of regulatory elements (McGaughey et al., 2008) and there have been many examples across multiple species of enhancer activity being conserved without obvious sequence conservation (Fisher et al., 2006; Yang et al., 2015; Ludwig et al., 2000; Hare et al., 2008). As a result, there has been considerable interest in predicting enhancer function from sequence (Chen et al., 2017), to provide a deeper understanding of how mutation at the sequence level can lead to compensation and maintenance of robust enhancer functionality (Khoueiry et al., 2017; Duque et al., 2014). Most of these studies have focused on well characterized systems in which either the key TFs or the architecture of relevant CREs is already known thereby limiting their use in a broader context.

In an attempt to bypass the need for either sequence conservation or locus specific knowledge, some recent studies have focused on understanding the genome-wide effects of CRE evolution using epigenetic marks which are thought to be proxies for regulatory activity. Recent work on DNA methylation (Qu et al., 2018) used a phylogeny-aware approach to detect gain and loss of methylation, reporting an expansion of methylation in mammals with lineage specific properties. Other studies which have focused on characterizing CREs with histone modifications (Villar et al., 2015) or nascent RNAs (Danko et al., 2018) in mammals and primates, used heuristic methods to report higher turnover rates in enhancers than promoters, as well as a wide variety of properties related to constraint on regulatory architectures. In addition to work characterizing the static landscape of regulatory activity some work has considered the conservation of dynamic responses in immune induction across the primates and reported relatively high conservation (Danko et al., 2018). However, it remains unclear how to create a general and robust, yet reasonably simple modeling approach for analyzing comparative epigenetic data.

In this work, we develop a new framework for analyzing comparative epigenomic data and apply it to a subset of histone modification data from Villar et al. (2015). We report on variety of results comparing the evolution of enhancers, promoter, and between subsets of promoters and enhancers based on characteristics of their associated genes. We also report on potential causes for discordance between sequence and functional constraint.

3.2 Results

3.2.1 Inferring evolutionary dynamics of epigenetic marks with a phyloHMM

We developed a phylogenetic hidden Markov model, epiPhyloHMM, to reconstruct the evolutionary histories of enhancers and promoters. PhyloHMMs are hidden Markov models whose hidden states define distinct phylogenetic models drawn from a finite set, allowing them to jointly consider how substitutions (and gains/losses) occur along branches of the phylogeny, and adjacent genomic sites (Siepel and Haussler, 2005; Felsenstein and Churchill, 1996). The epiPhyloHMM model consists of two components: a mixture negative binomial emission model for ChIP-seq peak calling, and a transition model with probabilities defined by an evolutionary process. Using a phyloHMM to model multi-species epigenetic data helps address issues caused by gaps in the alignment and the noisiness from variable quality ChIP-seq data.

The core idea behind epiPhyloHMM is that we can model evolutionary gain

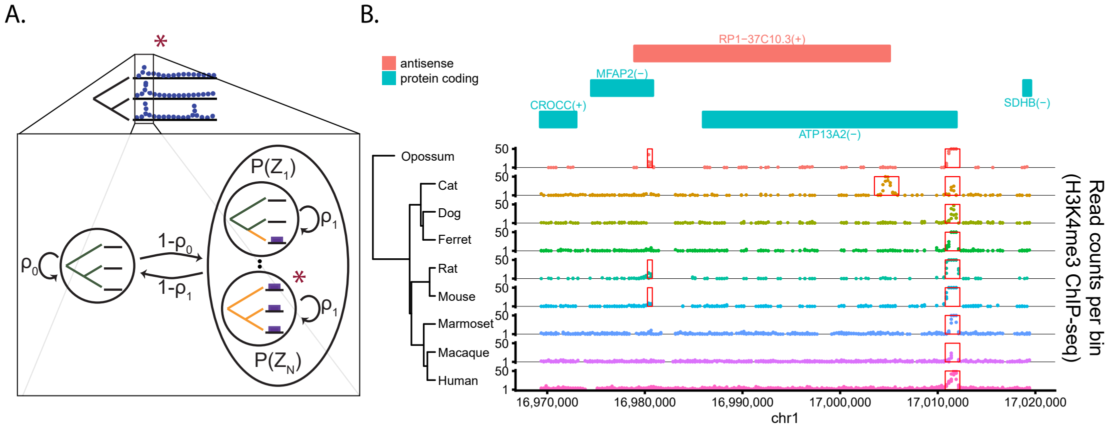


Figure 3.1: **The epiPhylo model** (A) State transition diagram of the epiPhylo model, illustrating the sparse transition between the state with no elements present in any species and any single state with an element present in one or more species. The red star indicates the selected state for the focal site illustrated in the cartoon data. (B) Example of state calls on real H3K4me3 ChIP-seq data on a subset of species from Villar et al. (2015).

& loss of epigenomic marks in a manner similar to the classic DNA substitution models, but handle the spatial distribution of epigenetic marks with an HMM. This is accomplished by labeling the tips of a phylogenetic tree using a two character 0/1 (absent/present) model to construct a state space (fig. 3.1) for the HMM. The probability of each state is then computed using the stationary probability of the characters on the tree π_1 ($\pi_1 = 1 - \pi_0$), and the turnover rate γ. In our model γ is the instantaneous rate of a gain or loss under a continuous time Markov process. To prevent the state space from being exponential (2^N) in the number of species, and thus intractable for explicit calculations using the HMM, we enumerated only a subset of tip labelings on the tree based on scenarios that could occur with three or fewer mutations on the tree.

We used epiPhyloHMM to analyze ChIP-seq data, so the noise model for raw read counts given 0/1 characters at the tips of the tree was a mixture negative binomial distribution, which has become widely used for modeling the variance in sequencing count data (Love et al., 2014; Anders et al., 2013). For the 0 state we used a single component mixture model and three components to handle peaks of different heights for the 1 state. A separate emissions model was trained in each species using data aligned to a subset of its native genome. Once the data was lifted over to a central reference genome, the emission model was augmented with a scale parameter ζ, to accommodate alignment gaps between species which result in missing data in the frame of the reference genome. The evolutionary parameters π_0, γ, and the state autocorrelation parameter ρ_0, are then fit using the data mapped to the central reference genome (see Methods).

3.2.2 Simulation Study

To test the power of our model to recover the true peak calls and parameters, we simulated data under the model used for inference, allowing for all 2^N states, given a variety of genome sizes, phylogenetic trees, and values of the turnover parameter (γ). We fixed the parameters for the peak calling model to their true values to isolate the ability of the model to correctly estimate the evolutionary parameters of interest.

With regard to calling peaks at the species level, we found that epiPhylo performed well by both precision and recall metrics but suffered slightly at higher turnover rates (fig. 3.2). Model predictions were evaluated on a per-bin basis so that a prediction was considered a true prediction if the model's state call for a bin in a particular species matched that bin's state in the simulation. Each bin in a multi-bin element was evaluated separately. Across a broad range of rate parameters and tree topologies, estimates of γ converged to an inflated, but stable

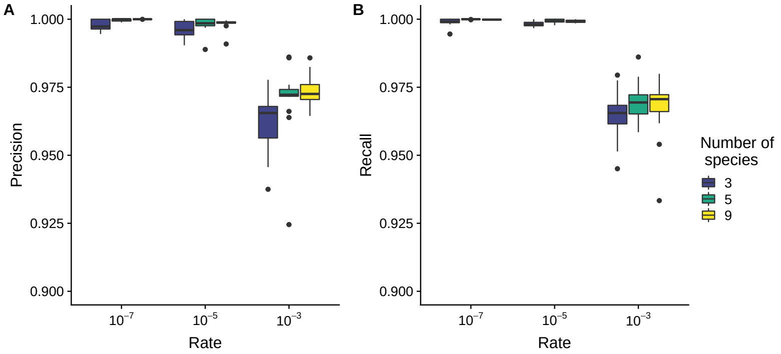


Figure 3.2: **EpiPhylo performance as a function of rate for simulated data.** (A) Precision of the epiPhylo model based on the number of 250bp bins for each species with correct state calls across three different trees. (B) Recall of the epiPhylo model based on the number of 250bp bins for each species with correct state calls across three different trees. For both calculations, the state that epiPhylo calls is used to assign active elements across species which are then compared to the true peaks at the species level. All possible species configurations were simulated, and the epiPhylo model was fit with all configurations that required three or fewer mutations.

estimate of γ with as little as 250KB when γ was large and as much as 250MB when γ was small (fig. 3.7). The stationary probability for an absent element π_1 was systematically underestimated while estimates of the autocorrelation ρ_1, converged to the true values given as little as 2.5MB of data (fig. 3.8-3.9). Systematic biases in the fitted values of γ and π_1 occur to accommodate due to a ridge in the likelihood surface (fig. 3.10), however this does not greatly impact the recovery of true regulatory elements in our simulation.

3.2.3 Decoupling the HMM and phylogenetic models for rigorous hypothesis testing and ancestral reconstruction

Next we became interested in using an evolutionary framework for comparing turnover rates between groups of elements with different annotations. While phyloHMMs are powerful for segmenting the genome, the cost of fitting more parameters with epiPhyloHMM becomes expensive, and the bias in the parameter estimates prevents recovery of the true parameters. To address these issues we adapted the standard phylogenetic model to accommodate probablistic state labels on the tips of the tree. Then we used the genomic segmentations provided by epiPhyloHMM to create a single epigenetic state in each species per contiguous epiPhylo element. First we grouped adjacent bins with the same state, then computed a joint likelihood across all bins conditioned first on the absent state, then on the present state. The procedure to test for different rates of turnover between subgroups of elements is as follows: (1) group epigenetic elements together by one or more annotations and fit a shared rate across all groups with all other parameters fit separately for each annotation to represent the null hypothesis of a uniform evolutionary rate across the annotated groups, and (2) as the alternative hypothesis, we can refit the models with separate rates for each annotation $a \in A$, then (3) compare the models using a likelihood ratio test. The alternative models are then nested within the null model by breaking the symmetry between rate parameters (γ) for all groups of elements and adding $|A| - 1$ degrees of freedom. This provides an efficient and rigorous method to detect differences in turnover rate for disjoint groups of epigenetic marks with one additional degree of freedom.

Another advantage of using a phylogenetic approach for analyzing epigenetic data over previous heuristic approaches is the ability to identify gain/loss events that occur on non-terminal branches of the tree. To accommodate the probabilistic allele labels computed by epiPhylo, we applied the previously described altered phylogenetic model to accommodate probabilistic allele labels at the tips. To accommodate potential differences in phylogenetic parameters

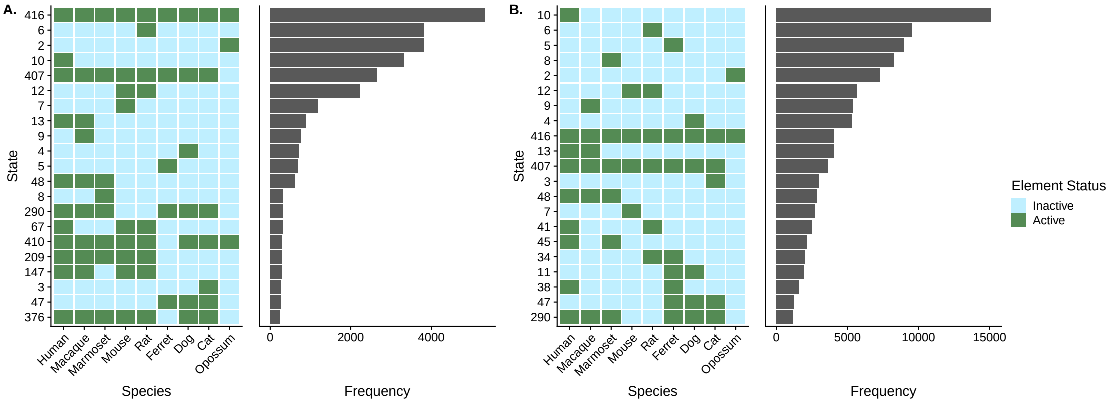


Figure 3.3: **Distribution of state calls by epiPhyloHMM.** (A) The distribution of state calls from epiPhyloHMM for the H3K4me3 mark. (B) The distribution of state calls from epiPhyloHMM for the H3K27Ac mark.

between enhancers and promoters we partitioned the sites with an H3K27Ac mark in one or more species into enhancers and promoters based on proximity to transcriptional start sites (TSS) in humans (fig. 3.11). We note that due to the human centric nature of the alignment framework, we tend to observe increasing numbers of losses on branches with a longer time to most recent common ancestor (TMRCA) with human (fig. 3.12). This provides that caveat that lineage specific rate tests may be unreliable if not constructed carefully.

3.2.4 Promoters show deeper epigenetic conservation than enhancers

We ran epiPhyloHMM on previously published comparative H3K4me3 and H3K27Ac ChIP-seq data from placental mammals, producing a genome wide segmentation. This analysis produced an average of ~16,000 and ~47,000 elements per species for the H3K4me3 and H3K27Ac mark respectively (fig. 3.13). The excess of H3K27Ac sites was in line with previous findings as the H3K27Ac mark covers both enhancers and promoters while the H3K4me3 mark is more specific to promoters. The distributions of state assignments were also highly distinct, with the fully conserved state being the most common for the H3K4me3 mark while being the ninth most common state for the H3K27Ac marks, with a much lower frequency than most of the single species states. The observation of lower rates of turnover in enhancers than promoters is in line with previous literature suggesting that promoters are more deeply conserved than enhancers (Villar et al., 2015).

To more directly investigate the differences in turnover rate between enhancers and promoters we partitioned the sites with an H3K27Ac mark based on proximity to transcriptional start sites (TSS) in humans as previously described (fig. 3.11). We then applied our likelihood ratio framework to derive rigorous estimates of turnover rate in each group and tested directly for a difference in rate using a tree based on real divergence times. We estimated a turnover rate of roughly 0.0075 events/mya for enhancers and 0.0035 events/mya for promoters (fig. 3.4A). We also estimated the half-life for enhancers and promoters and found that the half life for promoters was 424mya while the half life for enhancers was 302mya (see Methods). The enhancer half-life estimate is similar to previous estimate from Villar et al. (2015) of 296mya, however the promoter estimate is considerably shorter than the previous estimate of 939mya. The difference in the promoter half-life estimates may be partially explained by the epiPhyloHMM model being unable to fully accommodate the noisy nature of the H3K27Ac mark, resulting in an excess of gain/loss events thus increasing the estimated turnover rate. However, we do not see a similar effect on our estimate of enhancer half-life so we cannot easily select one estimate as superior to the other.

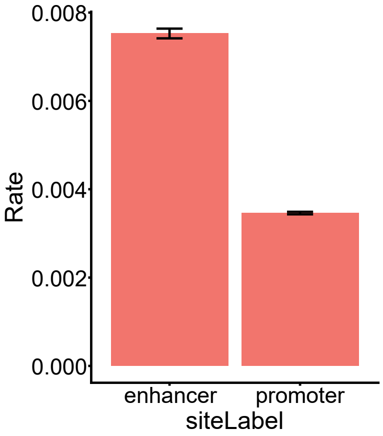


Figure 3.4: **Promoters show greater epigenetic conservation than enhancers**. Rates of turnover inferred for enhancers and promoters using the H3K27Ac mark on an ultrametric tree with branches scaled by chronological divergence times (Kumar et al., 2017). Error bars correspond to the maximum likelihood rate estimate +/− one standard error. The enhancer turnover rate is roughly twice as high as the promoter rate ($p \ll 10^{-300}$ via likelihood ratio test).

3.2.5 Pleiotropy and mutational intolerance of associated genes correlate with CRE turnover

Next, we investigated how tissue specificity of expression and sequence constraint of genes affected the turnover of putatively linked CREs. Each enhancer and promoter was assigned to a single gene based on a combination of distance based rules. Genes that were in the neighborhood of more than one gene were excluded from this analysis (fig. 3.11).

First we sought to determine whether stability of gene expression across tissues correlated with promoter turnover rates. Based on GTEx data, we annotated genes as being housekeeping, intermediate, or variable, based on their expression variance across tissues. Then, using the H3K4me3 mark, we compared the turnover rates of promoters for mean-expression-matched variable and housekeeping genes, and found that the promoters for housekeeping genes turned over at roughly half the rate of genes with variable expression (fig. 3.5A). This result agrees with previous literature and is consistent with the idea that highly pleiotropic loci are under stronger constraint than non-pleiotropic loci (Villar et al., 2015). Notably we do not see any difference in the turnover rate of putatively linked enhancers (H3K27Ac) of housekeeping and variable genes (fig .3.15). One possible explanation for this observation is that most housekeeping genes are primarily regulated by their promoters and any enhancers they do have tend to be close to the promoter and thus are filtered out or mis-annotated by our pipeline (Zabidi et al., 2015).

Next we investigated whether the promoter conservation across species was related to gene intolerance for ultra-rare loss of function (LoF) mutations. Since ultra-rare variants are almost always only present in single copies, genic intolerance for ultra-rare mutation is a proxy for haploinsufficiency. We compared fraction of elements associated with heterozygous LoF intolerant genes (pLI ≥ 0.9; Samocha et al. (2014)) to the number of species which had an active element for both promoters and enhancers (H3K27Ac) based on epiPhylo. We found that for both enhancers and promoters, being present in more species increased the chance that the associated gene is intolerant to heterozygous LoF mutations, although the effect was stronger in promoters than enhancers (fig. 3.5A-B). Together, these results indicate that the turnover rate of both enhancers and promoters decrease as selection on the sequence and pleiotropy of associated genes increases.

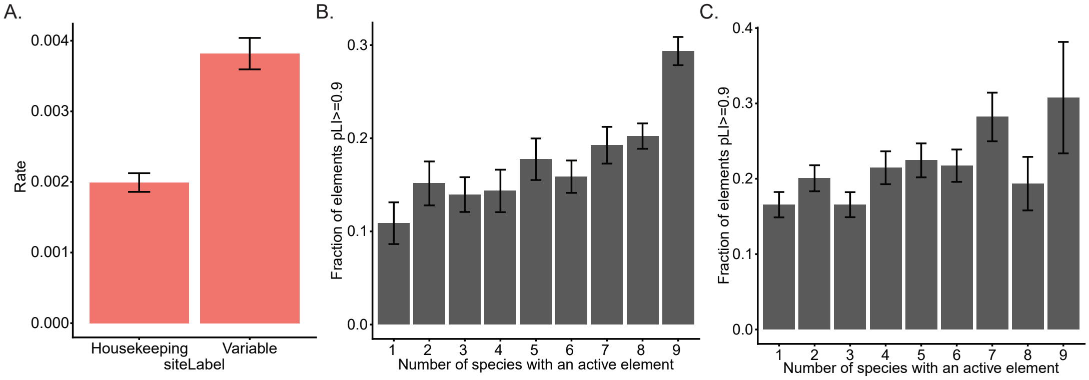


Figure 3.5: **Tissue specificity of gene expression and selection on coding sequence correlate with conservation of cis-regulatory elements.** (A) Estimated rates for promoter turnover based on the H3K4me3 mark where associated genes are classified as being housekeeping or tissue specific and matched for average expression ($p \ll 10^{-30}$). (B) Fraction of haploinsufficient genes based on genewise pLI scores for promoters with the H3K27Ac mark. pLI scores are a measure of intolerance to rare mutations. A pLI score ≥ 0.9 suggests that a gene is haploinsufficient. The probability of a gene having a pLI score ≥ 0.9 increases by 1.7% for each additional species the mark is conserved in. (C) Fraction of proximal enhancers with H3K27Ac mark associated with a haploinsufficient gene (pLI ≥ 0.9). The probability of a gene having a pLI score ≥ 0.9 increases by 1.3% for each additional species the mark is conserved in. Error bars for (B) and (C) represent the mean +/- one standard error.

3.2.6 Dosage sensitivity causes divergence between sequence and epigenetic conservation at promoters

Next we sought to investigate directly whether gene dominance affects sequence and epigenetic conservation. To get a list of genes with known dominance effects, we used a previously curated list of disease genes that were classified as having either a dominant or recessive mode of action (Berg et al., 2013; Blekhman et al., 2008). First we tested whether all disease genes showed a lower rate of turnover at promoters (H3K27Ac) than non-disease genes and found a roughly 25% decrease in rate (fig. 3.6A). We tested whether dominance of disease genes affected turnover rates of H3K27Ac marks at promoters and found that it did not (fig. 3.6B), likely because loss of promoter severely decreases, if not abolishes gene expression which is deleterious in a disease gene. In this line of reasoning we then theorized that while disease genes are equally intolerant of whole CRE turnover regardless of dominance status, recessive gene should tolerate smaller changes of expression while dominant one may not. This is because dominance status can be thought of as an indicator for dosage sensitivity as mutations with a recessive mode of action are likely insensitive to halving the expression of the wild type allele while dominant mutations are likely sensitive. Therefore CREs of genes that are highly dosage sensitive should exhibit stronger sequence conservation than those that are not, even when controlling for epigenetic conservation. To assess the strength of selection directly on the sequence we took the mean 100way-phastCons scores in the same sets of disease gene associated elements. We found that the promoter sequences of genes that were highly dosage sensitive (had a dominant mode of action), were under stronger constraint than those that were not (fig. 3.6C).

Another possible explanation for the observed difference in sequence conservation between CREs of dominant and recessive gene is different densities of regulatory sequence. To rule out this possibility we used putative TFBS annotated by the ENSEMBL regulatory build to compare the density of TFBS between the promoters of disease genes with recessive and dominant mechanisms, and found no difference (fig. 3.6D). Despite similar TFBS densities, we observed a decreased density eQTLs in the promoters of dominant disease genes relative to recessive ones, indicating a depletion of regulatory variants of sufficient effect size for detection (fig. 3.16). Together, these two lines of evidence suggest that selection on gene regulation occurs differently at multiple levels, dependent on the mechanism that mediates the deleterious phenotype.

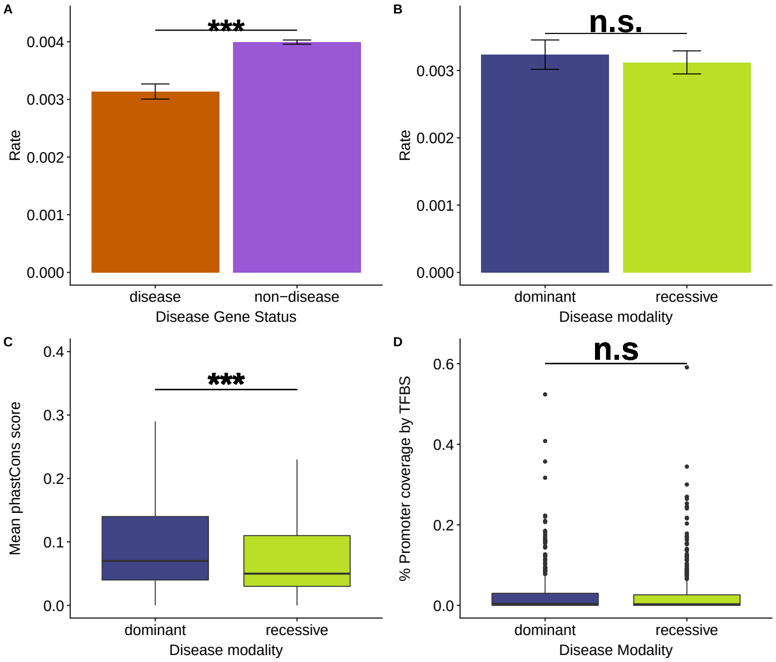


Figure 3.6: **Promoters of disease associated genes show decreased turnover rates, but the effect of dominance can only be seen at the sequence, not the epigenetic level.** (A) Comparison of promoter turnover rate between disease and non-disease associated genes. Promoters are based on the H3K27ac mark ($p \ll 10^{-30}$). (B) Comparison of promoter turnover rates (H3K27ac) between disease genes with dominant vs. recessive mode of action ($p = 0.61$). (C) Comparison of sequence conservation in promoter elements between disease genes with dominant vs. recessive mode of action ($p = 7.9 \cdot 10^{-8}$). (D) Percent of promoter nucleotides covered by TFBS from the ENSEMBL regulatory build for disease genes with dominant vs. recessive mode of action.

We then sought to test our hypothesis that differential dosage sensitivity leads to different patterns of sequence and epigenetic conservation on an orthogonally defined group of genes. Two groups that have been used a model for differential dosage sensitivity are genes involved in metabolism (low dosage sensitivity) and transcriptional regulation (high dosage sensitivity)(Wilkie, 1994; Veitia et al., 2018). We compared the epigenetic turnover rate of promoters marked with H3K27Ac and found that it turned over slightly more rapidly for metabolism than transcription genes. This observation suggests that, unlike with the dominant/recessive disease genes, there are differences in selection on the whole molecular function, therefore dominance is not the only differential selection relevant factor(fig. 3.17A). However, once we control for the number of turnover event at a given site as a proxy for epigenetic constraint, we see greater selection on the sequence in the promoters of transcription genes than metabolic genes, in line with expectations of greater dosage sensitivity for transcription genes (fig. 3.17B). Together, these results suggest that dominance is one of the factors driving differences between levels of sequence and epigenetic conservation.

3.3 Discussion

Changes in the activity of cis-regulatory elements which alter gene expression have been shown to drive phenotypic divergence between species (Prescott et al., 2015; Wray, 2007). Understanding the observed patterns of turnover and constraint of CREs is therefore informative for understanding the broader principles that govern the evolution of gene expression. The development of techniques such as ChIP-seq, STARR-seq, and various forms of BS-seq, for measuring evidence of regulatory function at the whole-genome level have opened up new opportunities for understanding the evolution of regulatory elements (Marinov and Kundaje, 2018; Arnold et al., 2014; Pai et al., 2011). Previous works have performed heuristic analysis of CRE evolution and gene expression, finding a positive correlation between the stability of gene expression and CRE conservation, as well as evidence of enhancer emergence correlating with adaptation (Berthelot et al., 2018; Villar et al., 2015). There have also been some model based approaches, focusing on integrating phylogenetic information for a variety of traits including replication timing and methylation (Yang et al., 2018;

Qu et al., 2018).

In our work we demonstrate a general framework for inferring relative levels of constraint on regulatory elements using a probabilistic model that integrates uncertainty from the raw data and alignments, then, by pooling information across regulatory elements, detects differences in turnover rate. By applying this framework we recapitulate previous work showing that enhancers turnover more rapidly than promoters (Villar et al., 2015), although we derive shorter half-life estimates, probably partially due to the noisy nature of the data. We also found that stability of gene expression across tissues correlates with conservation across species for promoters but not enhancers. The positive correlation between promoter conservation and gene expression is in agreement with previous work which found that conservation of gene expression was elevated in housekeeping genes (Berthelot et al., 2018). There are several possible technical reasons for the difference between enhancers and genes (e.g. imperfect enhancer-gene assignment, and noisy data) due to housekeeping genes having a fundamentally different regulatory architecture than tissue specific genes. For example, previous work has suggested that enhancers of house-keeping gene are much more likely to be in the core promoter region or the 5′ UTR, characteristics that could result in them being labeled promoters under our rules (Zabidi et al., 2015).

We also find several interesting results from the comparison of sequence and functional conservation. Our two key results were that (1) promoter and proximal enhancer conservation is higher for haploinsufficient genes and (2) even among essential genes, haploinsufficient genes show elevated levels of sequence conservation after controlling for epigenetic conservation. We propose that this result is due to sequence and epigenetic conservation being indicators of fundamentally different types of constraint. While no essential gene can afford to lose its promoter (indicated by loss of the epigenetic mark), and thus expression, it has been well documented that some genes which are less dosage-sensitive can tolerate more potentially expression altering sequence mutations (Petrovski et al., 2015). Thus this divergence between epigenetic and sequence constraint is potentially informative about the mode of selection at that locus.

There are a number of caveats for our work. First, all data is aligned to the human reference genome and uses human annotations for all analysis. This strategy creates a reference bias and makes lineage specific tests difficult to perform. This limitation may be overcome by using a meta-genome that represents all genomes equally well, for example by using a slight variant of the HAL format (Hickey et al., 2013). More broadly, the quality of cross-species alignments and the systematic biases often present in them (e.g. alignments being better near genes), require careful consideration when performing analysis. There is also the noisiness of the data itself, which can seen in the large number of slightly offset peaks which clearly mark the same regulatory element. In our analysis we tried to address these concerns to some degree by only performing tests between elements that were marked with the same histone modification and therefore had similar noise properties, but this noise still almost certainly increased our estimates of turnover rate. Errors in the cross-species alignment and the data combine to make analysis especially difficult in the HMM framework, which requires data features be well aligned. Although we have tried to address both of these problems they remain in need of further work.

Second, all of our mark annotation and gene assignments are based on distance rules with respect to the human TSS annotations. In addition to the previously discussed reference bias, its is also likely that a significant number of enhancers (and to a lesser degree promoters) are mis-assigned. Experimental work to link enhancers to the correct genes, either via 3D-chromatin capture (Sanyal et al., 2012; Jin et al., 2013; Mifsud et al., 2015), or large scale genome editing (Fulco et al., 2016), may address some of these issues although the requirement to perform such assays across multiple species in well matched tissue samples may prove difficult.

Third, although histone marks such as H3K27Ac and H3K4me3 are strongly associated with genomic function, they are noisy, imperfect measures of regulatory activity (Benton et al., 2017). Assays that measure genomic function with improved specificity and and cover a narrower span of DNA should provide better quantification of CRE evolution and make it easier to link sequence and functional evolution.

In summation, this work presents a general framework for analyzing epigenetic traits, and applies that framework to learn features of CRE evolution and investigate the relationship between sequence and functional evolution. Our observations are largely qualitatively consistent with previous work and theory (Villar et al., 2015; Petrovski et al., 2015), and provide a way forward for studying regulatory evolution.

3.4 Acknowledgments

This research was supported in part by NIGMS grant R35GM127070. The content is solely the responsibility of the authors and does not necessarily represent

the official views of the US National Institutes of Health.

3.5 Methods

3.5.1 ChIP-seq Data Preparation

All ChIP-seq data obtained from Villar et al. (2015). Reads were then aligned to their native genomes (Mikkelsen et al., 2007; Consortium, 2002; Peng et al., 2014; Consortium et al., 2014; Lindblad-Toh et al., 2005, 2011; Consortium, 2004; Yan et al., 2011) (obtained from the UCSC genome browser) using bowtie2 (v 2.2.9) (Langmead and Salzberg, 2012). Each read is then collapsed to cover the single base at the center of the read and lifted over to central reference genome using liftOver (Hinrichs et al., 2006) (currently hg38). The lifted-over reads are then converted to a bigwig of read coverage. Finally, reads are summed into bins of 250bp. Regions that were not mappable in any other species besides human (hg38) were excluded from this analysis, leaving 2.97Gb remaining in the multiple alignment.

3.5.2 Peak Caller Model

A simpler peak calling HMM is used to pre-fit some values to simplify the latter fitting of the epiPhyloHMM model. The probability of state p (peak/no-peak) at bin j is a negative binomial mixture model computed as:

$$P(\tilde{\mathbf{x}}_j|\tilde{\mu}_p, \gamma_j, \tilde{\mathbf{w}}_p, \theta_D, \tilde{\mathbf{s}}) = \sum_m \left[w_m \cdot \prod_r nbinom\Big(x_{j,r}|\mu_{p,m}\gamma_j s_r, f(\mu_{p,m}\gamma_j s_r, \theta_D)\Big)\right]$$

where $\tilde{\mathbf{x}}_j$ is the read counts for each replicate. The mean for a given state p and mixture component m is $\mu_{p,m}$. Each component m is given weight w_m where $\sum_m w_m = 1$. Since some sites will not be mappable between genome we introduce γ_j which scales $u_{p,m}$ by the fraction of bases in bin j that are mappable between the native and central reference genome. To account for inter-library differences in sequencing depth we introduce s_r, which is calculated as:

$$s_r = \frac{1}{max(\vec{s})} \sum_j x_{j,r}$$

To account for documented differences in the dispersion of replicates for different mean depths, we subsample the genome to get roughly similar numbers of low, medium, and high coverage sites, then pre-estimate a mean-dispersion function using DESEQ2. We then use the parameters of that function θ_D, to compute the dispersion for all mixture components m and loci j.

The transition model is a simple two state model with auto-correlation ρ_A for the peak state and ρ_I for the background state.

3.5.3 epiPhyloHMM Model

The epiPhyloHMM model enumerates a finite set of states $\{s_1, ..., s_E\}$ at the tips of the tree based on a fixed maximum number of "mutation" events that partition the species tree. The emission probability for a given state s_e at site j is computed as $\prod_t P(x|s_{e,t})$, the joint probability of tip t in having the label specified by state s_e, for all tips t. More explicitly it is written as

$$P(\tilde{\mathbf{x}}_j|s_e, \tilde{\mu}, \gamma_j, \tilde{\mathbf{w}}, \theta_D, \tilde{\mathbf{s}}) = \prod_t \left(\prod_p \left[P(\tilde{\mathbf{x}}_{jt}|\tilde{\mu}_{pt}, \gamma_{jt}, \tilde{\mathbf{w}}_{pt}, \theta_D, \tilde{\mathbf{s}}_{\mathbf{t}})^{\delta(p=s_{e,t})} \right] \right)$$

However, where there is a large alignment gap in a species (set to 5Kb for overall model fitting), the probability of an active element in that region is forced to 0 for within that species, presuming a deletion.

The transition matrix is parameterized along the diagonal by ρ_I for the fully inactive state and ρ_A for all other states. The probability of entering any other state from the fully inactive state I is $(1 - \rho_I) \cdot P(z)$. $P(z)$ is the probability of state z computed by Felsenstein's algorithm given a set of "base" (peak/no-peak) frequencies π and mutation rates γ and normalized by $\sum_{z \neq I} P(z)$. The rate of transitions from any tree with one or more active elements to the fully inactive state is $(1 - \rho_A)$. All other elements of the transition matrix are 0, preventing the direct transition between any two different trees with active elements.

3.5.4 Model fitting

The epiPhyloHMM model is fit in several successive steps for efficiency. First, the peak calling models are fit separately for each species using the reads mapped to a subset of the native genome (125Mb) using the L-BFGS-B algorithm. The negative binomial mixture parameters of the peak calling model, w and μ, are then saved for use in the full epiPhyloHMM model.

The genome is then split into twenty similarly sized chunks, with chunk break-points determined by long regions where there is no alignment to the human genome by any other species. Bins with less than 15% of their sequence aligning to the human genome are treated as missing data on a per species basis. Then the ρ_A, ρ_I, π_0, and γ parameters in the epiPhyloHMM model are fit with the peak calling parameters held constant using L-BFGS-B. The resulted are then cleaned by masking out small element calls caused by peak fragmentation using the following heuristic.

1. Sets of elements that are only seperated by a single bin (a gap enforced by model sparsity) are grouped together.
2. A sum of scale factors for over all species for each element in the group is then computed
3. The element with the greatest alignability to the genome based on the sum of scale factors is kept
4. Additional elements are kept if their score exceeds a threshold t ($t = 16$).
5. Regions containing discarded elements are masked by setting their scale factors to 0.

The model then undergoes a second round of fitting post cleaning. The model is then re-run genome-wide using the median values of the ρ_A, ρ_I, π_0, and γ parameters to obtain a final set of calls.

3.5.5 Extracting and computing allele probabilities from epiPhylo

Histone mark elements were annotated based on the viterbi path from a fitted epiPhylo model. A single element probability was computed per allele by taking the product of the allelic probabilities from each bin.

3.5.6 Site annotation and gene association for enhancers and promoters

All distances were based on the Ensembl build 93 (Zerbino et al., 2018), accessed via BiomaRt (Durinck et al., 2005, 2009). Transcriptional start sites (TSS) for all transcripts were expanded by +/- 1.5kb and grouped by gene to create a promoter region. H3K4me3 elements that overlapped with only one promoter were annotated as promoter and associated with that gene. H3K4me3 elements that overlapped with more than one gene's promoter were annotated as an unassociated promoter (promoter_UA). H3K4me elements that did not overlap with any promoters were annotated as unknown (unk).

For H3K27Ac marks the same rules were used to label an elements as promoter or promoter_UA. For enhancer annotations the TSS were expanded by +/- 10kb. H3K27Ac elements that overlapped with only one expanded promoter were annotated as a proximal enhancer (enhancer_proximal) and associated with that gene. H3K27Ac elements that overlapped with more than one gene's expanded promoter were annotated as an unassociated proximal enhancer (enhancer_proximal_UA). H3K27Ac elements that did not overlap with any promoters but were still within 100Kb of a TSS were annotated as distal enhancers and associated with the closest gene (enhancer_distal). H3K27Ac elements that met none of these criteria were labeled as unknown. This scheme is represented in figure 3.11.

3.5.7 Ancestral reconstruction with probabalistic alleles and half life estimation

To perform ancestral reconstruction we implemented a standard phylogenetic model where the values for alleles at the tips of the tree were $P(x|\text{allele state})$ instead of the standard 0/1 encoding. We then performed the standard message passing algorithm on the phylogeny, then estimated the probabilities of a state transition on each branch as previously described in Siepel *et al.*. To make the values as comparable as possible, we used the H3K27Ac sites annotated as enhancers and promoters, leaving out the H3K3me3 sites. To estimate half life, estimated separate scaling factors for the enhancer and promoter trees, then estimated their respective half-lives adjusting for the stationary distribution as follows:

$$t_{\frac{1}{2}} = \frac{ln(2)}{\gamma \cdot \pi_P}$$

3.5.8 Enhancer vs. Promoter turnover analysis

All H3K27Ac elements annotated as any type of enhancer or promoter were fit with the stand alone phylogenetic model, first with a joint turnover rate, then separately. A p-value was computed using the LRT and confidence intervals are derived using the Fisher information matrix.

3.5.9 Housekeeping Vs. Variable gene analysis

Genes were annotated as house-keeping or variable using the GTEx v7 release (GTEx Consortium, 2017). The variance across tissues was computed for each gene then 20% with the smallest variance were annotated as housekeeping and the 20% largest variable were labeled as variable. The gene sets were then matched for mean expression across tissues using MatchIt (Ho et al., 2011). Then, using the H3K4me3 probabilistic alleles, shared and separate rates are fit for each set of elements on the Timetree and a p-value computed using the likelihood ratio test. Confidence intervals are derived using the Fisher information matrix.

3.5.10 pLI analysis

pLI values were obtained from, ExAC release 1 (`ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/manuscript_data`). The number of species which contain an active element is computed using their assigned state from the epiPhylo element. The effect of an additional species with an active element was estimated using linear regression. Error bars were computed using method of moments estimator for the standard error of p from the binomial distribution.

3.5.11 Dosage sensitivity gene conservation analysis

Lists of dominant and recessive genes were downloaded from the McArthur lab GitHub gene lists (`https://github.com/macarthur-lab`). These list are the union of the Berg and Blekhman dominant/recessive lists (Berg et al., 2013; Blekhman et al., 2008). Any genes that were in both lists were removed from the analysis. Phylogenetic models with both separate and shared rates were fit to compare both genes not in the list vs. genes in the list and dominant genes vs. recessive genes using the promoter associated H3K27Ac elements. H3K27Ac elements were mapped to grch37 using liftOver (Hinrichs et al., 2006) and mean 100way phastCons scores were calculated bwtool (Pohl and Beato, 2014). TFBS density was computed using the Ensembl regulatory build for grch37 on the lifted over elements (Zerbino et al., 2015).

Annotations of genes as "Metabolic" or "Generic Transcription Pathway" were taken from Reactome 2018 (Fabregat et al., 2016). Genes with only one of the functional annotations "R-HSA-1430728" or "R-HSA-212436" were selected for analysis. Both the phylogenetic rate and phastCons analysis were performed in the same manner as it was for the disease genes.

3.6 Supplemental figures

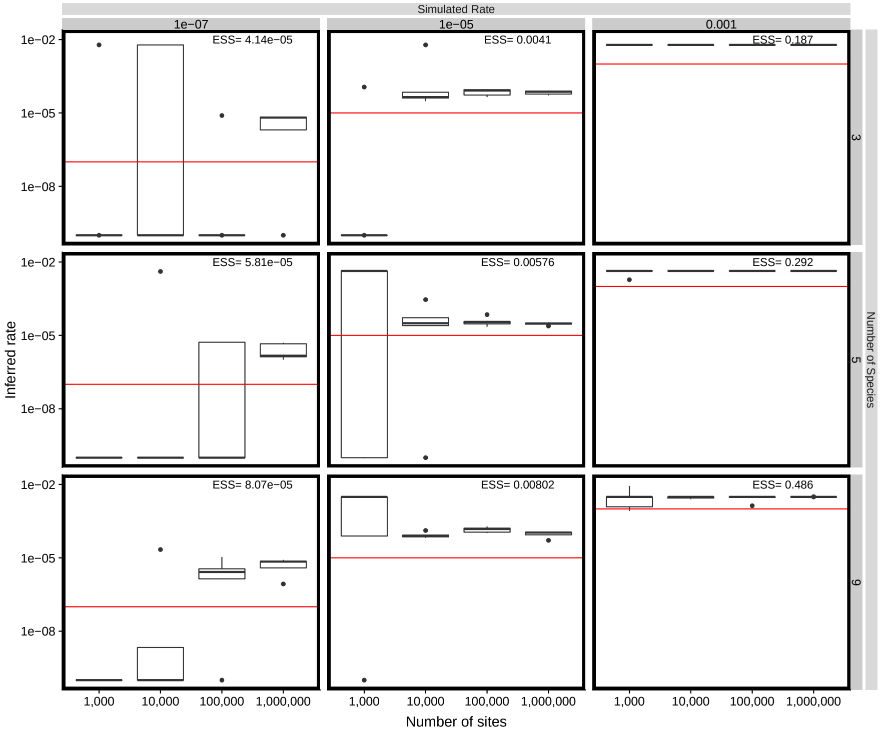


Figure 3.7: **Estimates of γ depend on the number of expected substitutions per site.** Raw count data was simulated under the epiPhyloHMM model, with all possible states being enumerated, for genomic regions of varying size (1,000 sites = 250KB) at varying rates for fixed ρ_A and π_A. epiPhyloHMM converges to inflated estimates of the correct values given increasing amounts of data, with convergence occurring more rapidly for scenarios with a greater number of expected substitutions per site.

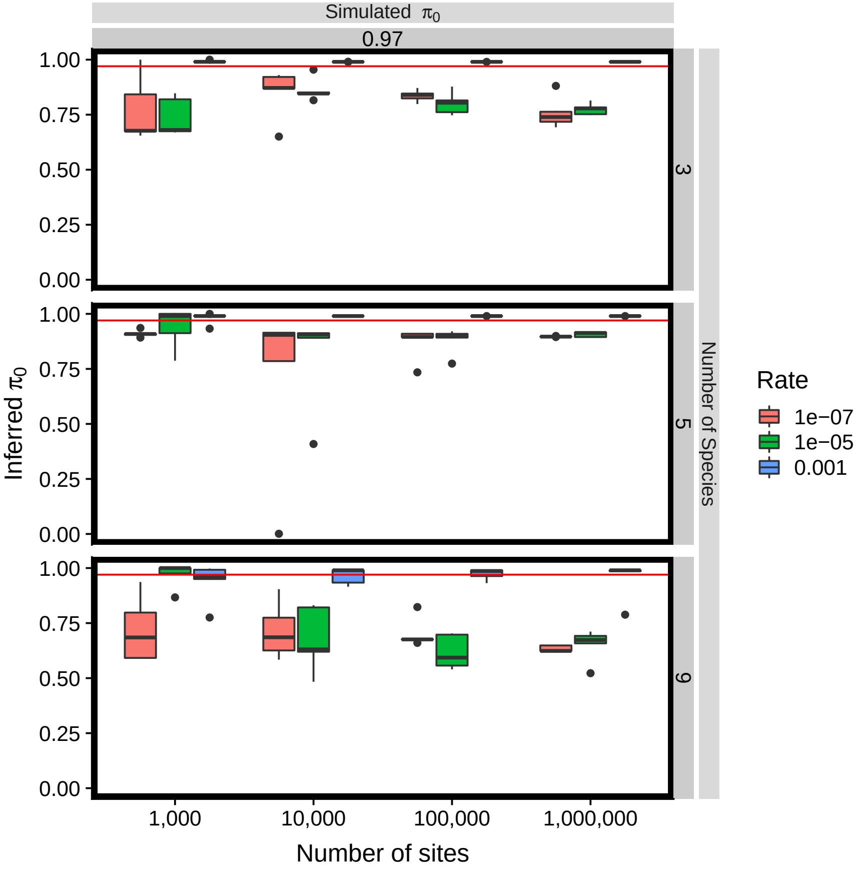


Figure 3.8: **Estimates of π_0 are biased by mis-estimates of rate.** Raw count data was simulated under the epiPhyloHMM model, with all possible states being enumerated, for genomic regions of varying size (1,000 sites = 250KB) at varying rates for fixed ρ_A and π_A. epiPhyloHMM converges to biased estimates of the correct values depending on the rate.

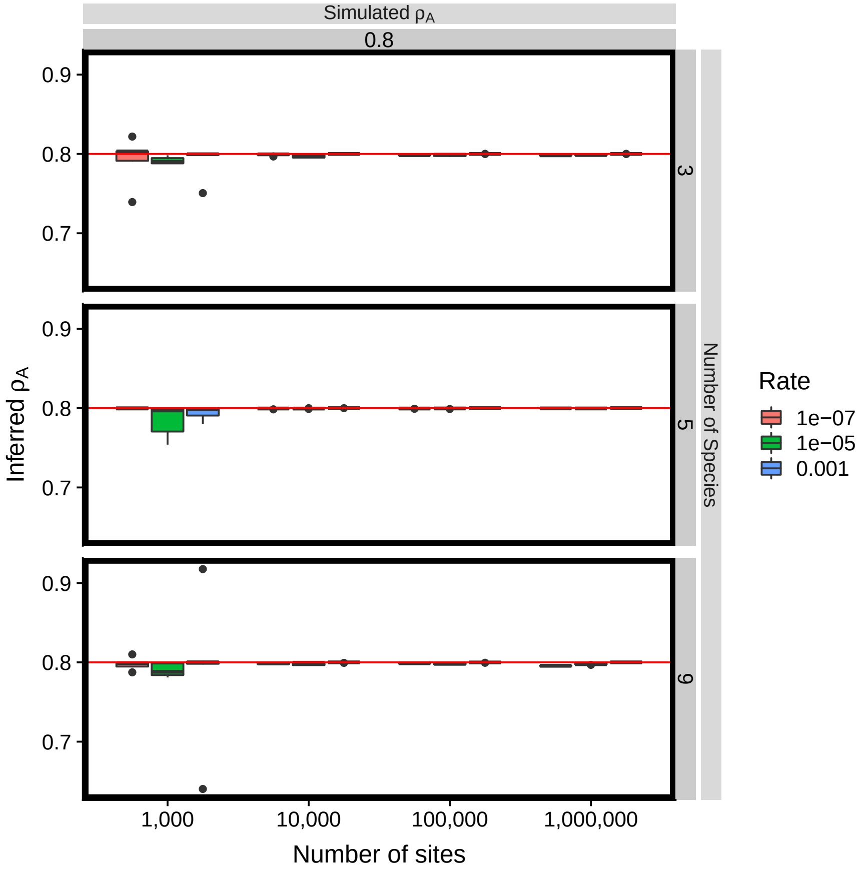


Figure 3.9: **Estimates of ρ_A converge rapidly to the true value** Raw count data was simulated under the epiPhyloHMM model, with all possible states being enumerated, for genomic regions of varying size (1,000 sites = 250KB) at varying rates for fixed ρ_A and π_A. epiPhyloHMM converges to correct estimates of the true value of ρ_A.

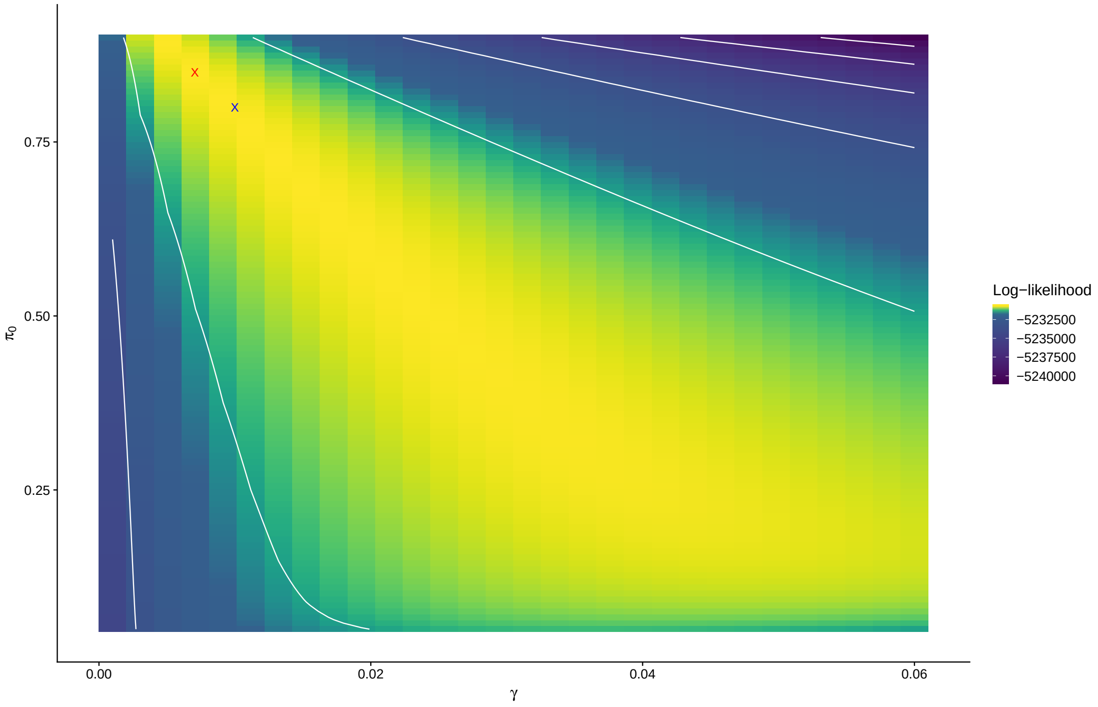


Figure 3.10: **Example of ridge in log-likelihood surface for fitting epiPhylo model.** Log-likelihood is computed on finite grid of γ and π_0 values with the auto-correlation parameters (ρ_0, ρ_1) fixed to the true values of the simulated data. The blue "X" indicates the true value while the red "X" indicates the MLE parameter values on the computed landscape.

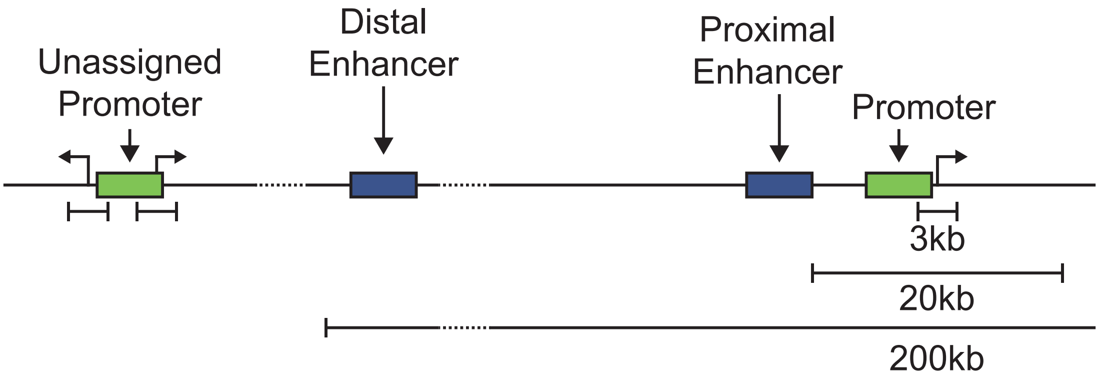


Figure 3.11: **Annotation scheme for epigenetic elements.** All genic distances and annotations are based on the human genome build GrCH38 (Zerbino et al., 2018). H3K4me3 elements were annotated as promoters if they were within +/ − 1.5kbp of a TSS. If they overlapped with TSS(s) for only one gene, they were associated with that gene. If they overlapped with TSSs from more than one gene, they were annotated as unassigned promoters. The same rules apply for annotating H3K27Ac marks as promoters, however there are additional rules for annotating them as enhancers. If a H3K27Ac element was within +/ − 10kbp but not within +/ − 1.5kbp of a TSS they were annotated as proximal enhancers and assigned analogously to promoters. If an H3K27Ac mark was between 10kbp and 100kbp away from the nearest TSS, they were annotated as distal enhancer and assigned to the gene of the nearest TSS. H3K27Ac elements further away than 100kbp were annotated as unknown.

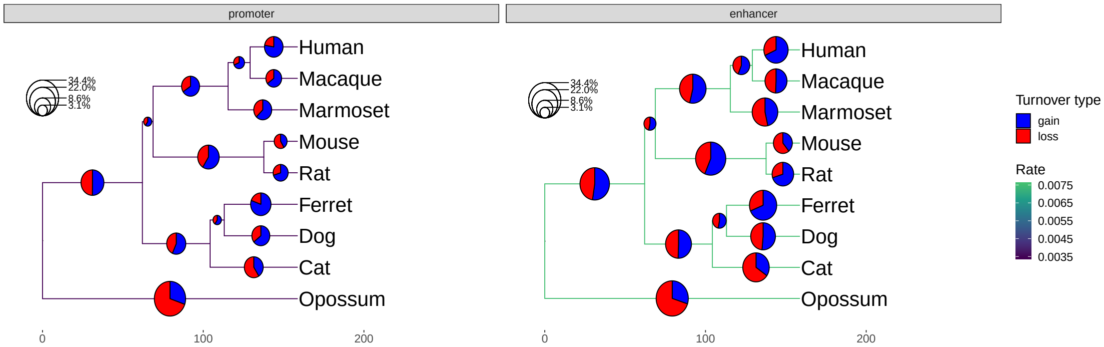


Figure 3.12: **Ancestral reconstruction of gain/loss events on a rooted tree.** Pie chart area per branch is proportional to the fraction of total enhancer/promoter state calls undergoing gain or loss. Numbers of gain/loss events were computed from pairwise marginals of the transition matrices on each branch. (A) Expected numbers of enhancer gains/losses based on H3K27Ac mark. (B) Expected numbers of promoter gains/losses based on H3K27Ac mark. Sites with an H3K27Ac mark in one or more species were partitioned into enhancers and promoters based on proximity to transcriptional start sites (TSS) in humans (fig. 3.11, see Methods).

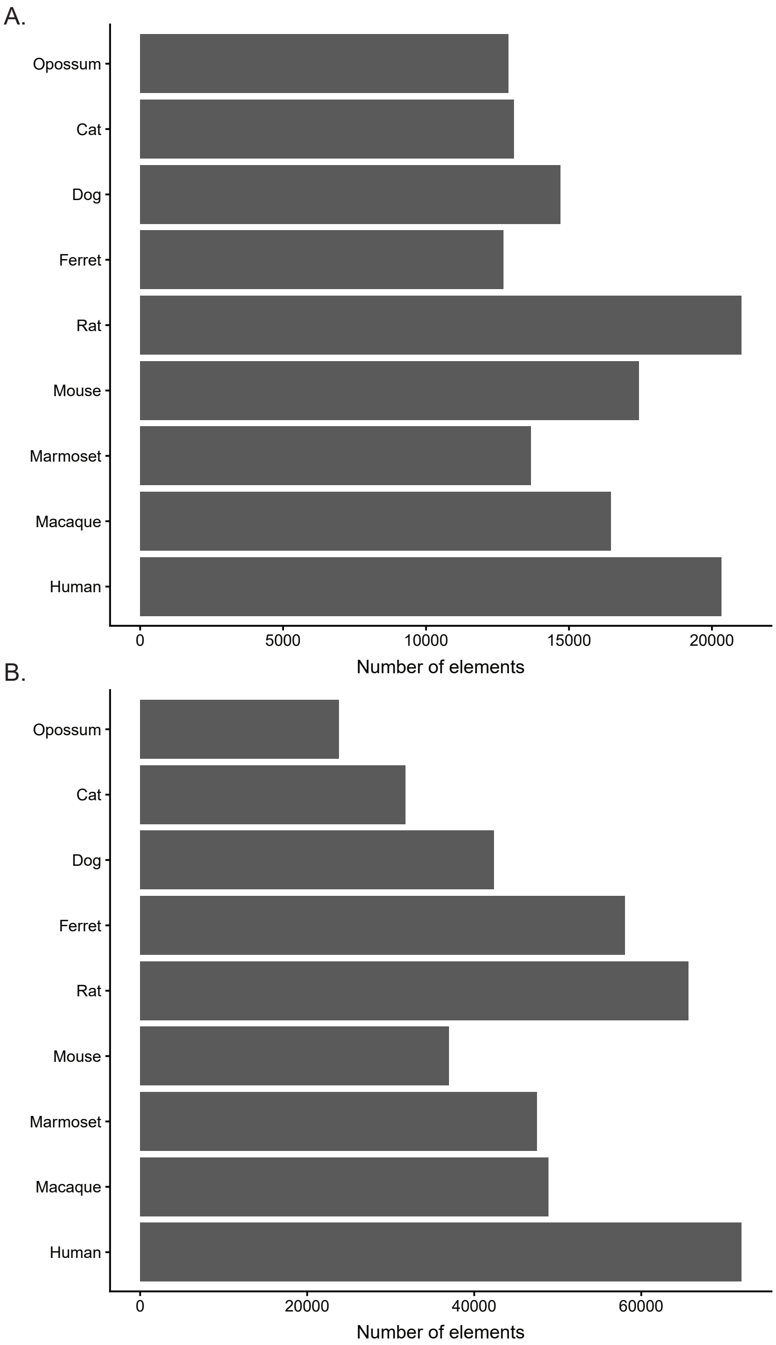


Figure 3.13: **Number of multi-species epigenetic state calls from epiPhylo based on viterbi decoding** (A) H3K4me3 (B) H3K27Ac

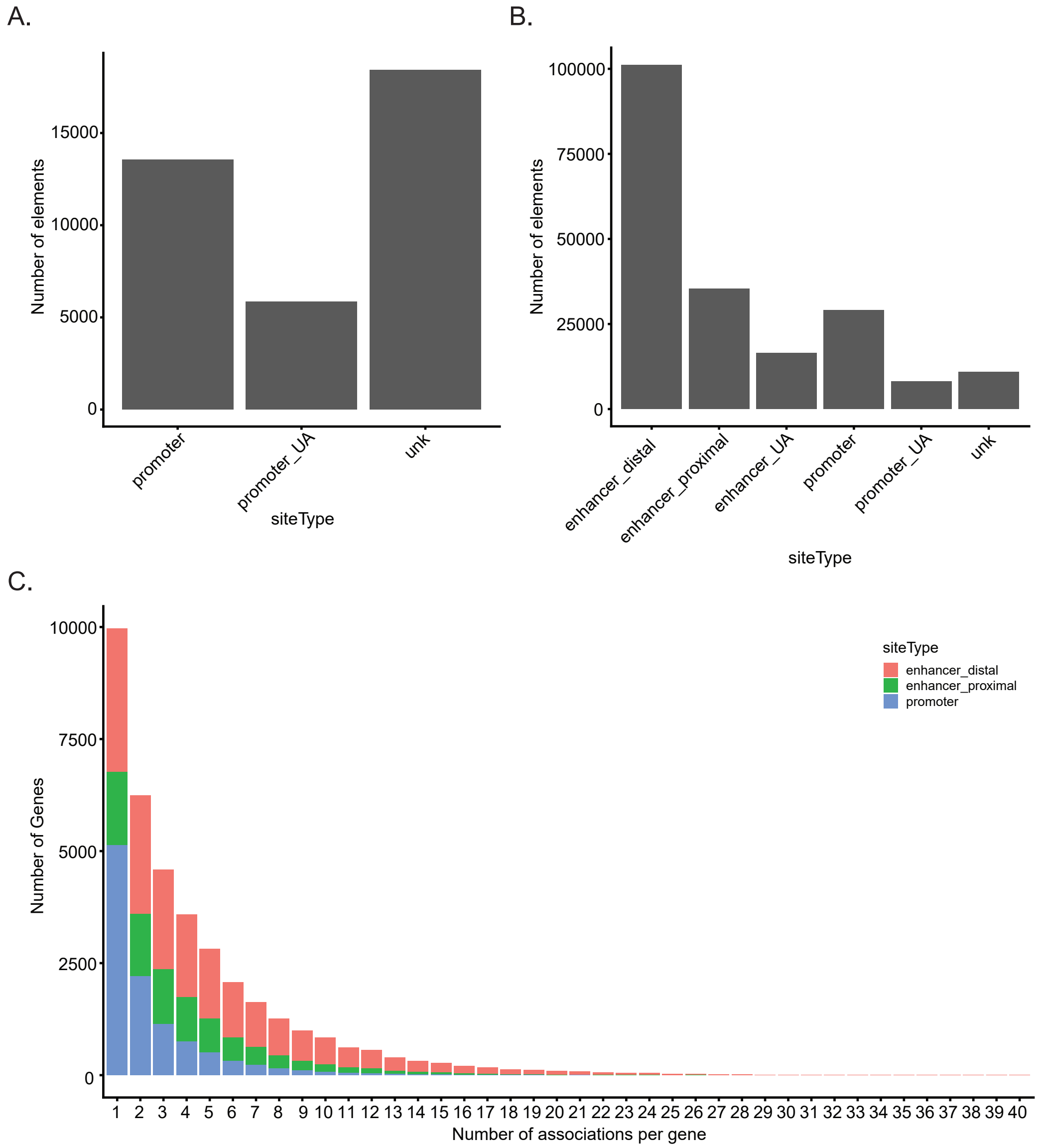


Figure 3.14: **Distribution of epigenetic element annotations.** (A) Distribution of annotations for elements with the H3K4me3 mark. (B) Distribution of annotations for elements with the H3K27Ac mark. (C) Distribution of the number of associations with H3K27Ac active elements per gene. The relative height of each of the color bars represents the fraction of associations that belong to each annotation category.

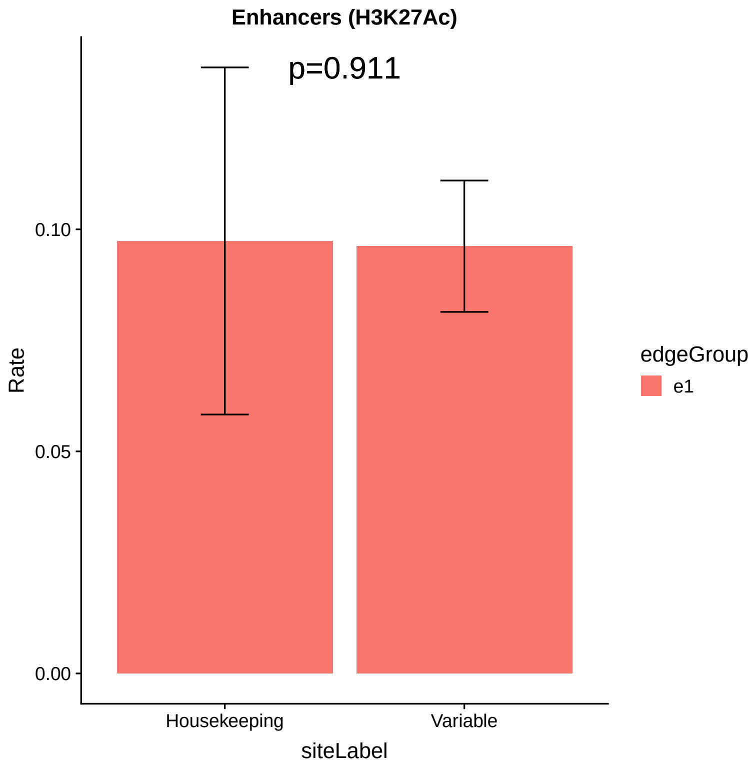


Figure 3.15: **No evidence for differential enhancer turnover rates for housekeeping vs. highly variably expressed genes**(A) Estimated rates for enhancer turnover based on the H3K27Ac mark where associated genes are classified as being housekeeping or tissue specific and matched for average expression.

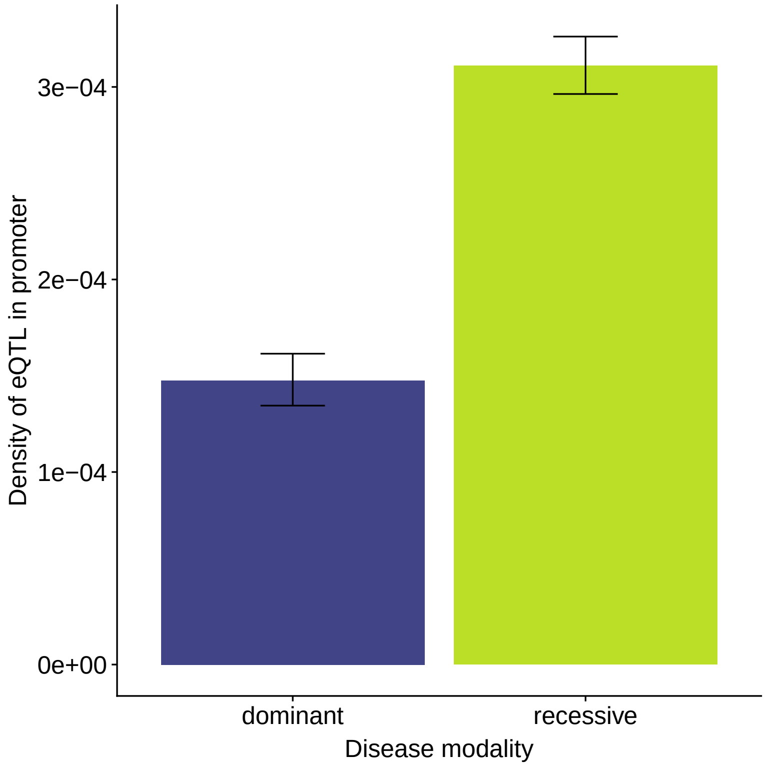


Figure 3.16: **eQTL density is depleted in dosage sensitive genes.** eQTL density based on GTEx (v7) significant eQTLs that intersect with epiPhylo elements.

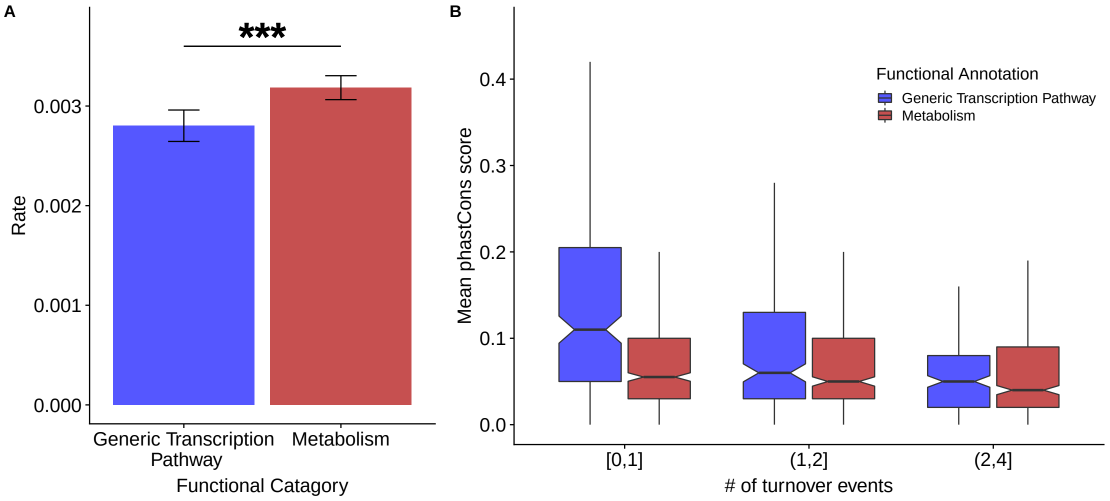


Figure 3.17: **Comparison of sequence and epigenetic turnover for promoters (H3K27Ac) associated with metabolism and transcription regulation genes.** (A) Estimated turnover rate for trancription pathway and metabolism genes ($p = 0.0074$). (B) Mean phastCons scores in promoters of trancription pathway and metabolism genes. Notches correspond to $1.58 \cdot IQR/\sqrt{n}$ corresponding to roughly a 95% confidence interval for the median.

CHAPTER 4

CONCLUDING PERSPECTIVE

The three chapters of this thesis represent distinct approaches to understanding the biology of cis-regulation of gene expression. Chapter one analyzes cis-regulation and gene expression from a global perspective, leveraging transcriptional data to posit specific TFs as dictating particular transcriptional programs without elucidating upon the mechanism at any individual loci. In contrast, chapter two focuses on understanding the quantitative architecture of a single cis-regulatory locus in a TF agnostic fashion, using gene editing data and statistical modeling to understand the contribution of individual enhancers to expression a specific gene. Instead of trying to understand how regulatory responses occur, chapter three seeks to understand the principles by which cis-regulatory loci change over time, modeling epigenetic data with phylogenetic models to compare turnover rates of different groups of CREs. This thesis leverages new experimental techniques to understand how CRE function relates to gene expression at both very short and very long timescales. By considering each perspective, this thesis improves the state of knowledge about CREs at both mechanistic and evolutionary levels by developing statistical methods for data from modern experimental techniques.

Despite the progress made by recent work, there are still many open questions which further unifying the mechanistic and evolutionary perspectives on CREs will allow the field to address. At present, it is still not well understood at a quantitative level how cis-regulatory architectures evolve. While other work (Villar et al., 2015; Berthelot et al., 2018; Arnold et al., 2014) and chapter three of this thesis look at wholesale gain/loss of elements, relatively little has been done to assess how the strength of both individual enhancers and whole CRE ensembles evolve in connection with the properties of their associated genes. Some open questions of interest are: (1) how rapidly does the quantitative strength of enhancers change between species? (2) under what conditions does synergy between enhancers evolve? and (3) how do properties of associated genes, like dosage sensitivity, shape the quantitative evolution of CREs? MPRA assays like STARR-seq (Arnold et al., 2013) could be used in combination with continuous trait models to address the first question. Detecting synergy would require a combination of gene editing approaches like those from Shin et al. (2016) and Hay et al. (2016), and high resolution chromatin conformation capture (Ma et al., 2017) to associate genes and CREs with candidate loci. Understanding the relationship of genic properties to CRE evolution could then be done using the data collected for the first two questions and devising statistical tests similar to those presented in chapter three.

A second area that warrants considerable investigation is the importance and prevalence of evolutionary compensation at CREs and the principles that govern it. Previous work has reported evidence of compensation via transcription factor binding site substitution both at individual loci (Hogues et al., 2008) and at a genomic scale in a small number of species (Khoueiry et al., 2017). Khoueiry et al. (2017) focused on enhancer sequence evolution as a function of intrinsic properties of the enhancer itself and found evidence for increased TFBS sequence turnover in TFBS dense enhancers, consistent with the TF collective model (Spitz and Furlong, 2012). At present I see two major avenues to further our understanding of evolutionary compensation: (1) investigating how dosage sensitivity and temporal control of associated genes shapes compensation of both whole the whole CRE and underlying sequence; and (2) developing methods for detecting compensation without relying upon knowledge of specific TFBS, thereby sidestepping the need for specific semi-mechanistic models of CREs. Addressing the first area to some degree is likely possible with existing datasets, either by comparing evidence of TFBS binding/CRE activity vs. sequence turnover for subgroups of genes in a manner similar to chapter three of this work. The second avenue might be addressed by exploiting machine learning methods which predict CRE function from sequence and using them to predict the activity trajectories of reconstructed ancestral sequences. These trajectories could then be compared to a neutral model to detect excesses of compensatory events.

Both of these areas of inquiry will become more pressing as genomewide quantitative measures of CRE activity become more common, thus exacerbating the need to move away from thinking of CREs using simple presence/absence models.

BIBLIOGRAPHY

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33**:831–838.

Allison, A. C., Cacabelos, R., Lombardi, V. R. M., Álvarez, X. A., and Vigo, C. (2001). Celastrol, a potent antioxidant and anti-inflammatory drug, as a possible treatment for Alzheimer's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **25**:1341–1357.

Amati, B., Brooks, M. W., Levy, N., Littlewood, T. D., Evan, G. I., and Land, H. (1993). Oncogenic activity of the c-Myc protein requires dimerization with Max. *Cell* **72**:233–245.

Amati, B., Dalton, S., Brooks, M. W., Littlewood, T. D., Evan, G. I., and Land, H. (1992). Transcriptional activation by the human c-Myc oncoprotein in yeast requires interaction with Max. *Nature* **359**:423–426.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protocols* **8**:1765–1786. 00015.

Andrulis, E. D., Guzmán, E., Döring, P., Werner, J., and Lis, J. T. (2000). High-resolution localization of Drosophila Spt5 and Spt6 at heat shock genes in vivo: Roles in promoter proximal pausing and transcription elongation. *Genes & Development* **14**:2635–2649.

Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., Lau, N. C., and Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nature Genetics* **46**:685–692.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, L. M., Rath, M., and Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**:1074–1077. 00034 PMID: 23328393.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**:729–740.

Benoist, C. and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promoter region. *Nature* **290**:304–310.

Benton, M. L., Talipineni, S. C., Kostka, D., and Capra, J. A. (2017). Genome-wide Enhancer Maps Differ Significantly in Genomic Distribution, Evolution, and Function. *bioRxiv* page 176610.

Berg, J. S., Adams, M., Nassar, N., Bizon, C., Lee, K., Schmitt, C. P., Wilhelmsen, K. C., and Evans, J. P. (2013). An informatics approach to analyzing the incidentalome. *Genet. Med.* **15**:36–44.

Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T., and Flicek, P. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution* **2**:152–163.

Bi, M., Naczki, C., Koritzinsky, M., Fels, D., Blais, J., Hu, N., Harding, H., Novoa, I., Varia, M., Raleigh, J., et al. (2005). ER stress-regulated translation increases tolerance to extreme hypoxia and promotes tumor growth. *The EMBO Journal* **24**:3470–3481.

Blekhman, R., Man, O., Herrmann, L., Boyko, A. R., Indap, A., Kosiol, C., Bustamante, C. D., Teshima, K. M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**:883–889.

Brown, M. S. and Goldstein, J. L. (1997). The SREBP Pathway: Regulation of Cholesterol Metabolism by Proteolysis of a Membrane-Bound Transcription Factor. *Cell* **89**:331–340.

Burdon, T., Sankaran, L., Wall, R. J., Spencer, M., and Hennighausen, L. (1991). Expression of a whey acidic protein transgene during mammary development. Evidence for different mechanisms of regulation during pregnancy and lactation. *J. Biol. Chem.* **266**:6909–6914.

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**:192–197.

Chen, L., Fish, A. E., and Capra, J. A. (2017). Sequence properties underlying gene regulatory enhancers are conserved across mammals. *bioRxiv* page 110676.

Churchman, L. S. and Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**:368–373.

Consortium, M. G. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**:520–562.

Consortium, R. G. S. P. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**:493–521.

Consortium, T. M. G. S. a. A., Worley, K. C., Warren, W. C., Rogers, J., Locke, D., Muzny, D. M., Mardis, E. R., Weinstock, G. M., Tardif, S. D., Aagaard, K. M., et al. (2014). The common marmoset genome provides insight into primate biology and evolution. *Nature Genetics* **46**:850–857.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., and Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics* **46**:1311–1320.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322**:1845–1848.

Crocker, J., Ilsley, G. R., and Stern, D. L. (2016). Quantitatively predictable control of *Drosophila* transcriptional enhancers *in vivo* with engineered transcription factors. *Nature Genetics* **48**:292–298.

Danko, C. G., Choate, L. A., Marks, B. A., Rice, E. J., Wang, Z., Chu, T., Martins, A. L., Dukler, N., Coonrod, S. A., Wojno, E. D. T., et al. (2018). Dynamic evolution of regulatory element ensembles in primate CD4 + T cells. *Nature Ecology & Evolution* page 1.

Danko, C. G., Hah, N., Luo, X., Martins, A. L., Core, L., Lis, J. T., Siepel, A., and Kraus, W. L. (2013). Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Molecular Cell* **50**:212–222.

Danko, C. G., Hyland, S. L., Core, L. J., Martins, A. L., Waters, C. T., Lee, H. W., Cheung, V. G., Kraus, W. L., Lis, J. T., and Siepel, A. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. *Nature Methods* [Online; accessed 2015-04-08].

de Nadal, E., Ammerer, G., and Posas, F. (2011). Controlling gene expression in response to stress. *Nat Rev Genet* **12**:833–845.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biol* **8**:e1000384.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21.

Dolken, L., Ruzsics, Z., Radle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., et al. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14**:1959–1972.

Dukler, N. (2018). *superEnhancerModelR: superEnhancerModelR*.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**:57–74.

Duque, T., Samee, M. A. H., Kazemian, M., Pham, H. N., Brodsky, M. H., and Sinha, S. (2014). Simulations of Enhancer Evolution Provide Mechanistic Insights into Gene Regulation. *Mol Biol Evol* **31**:184–200.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., Moor, B. D., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**:3439–3440.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4**:1184–1191.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2016). The Reactome pathway Knowledgebase. *Nucl. Acids Res.* **44**:D481–D487.

Felsenstein, J. and Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**:93–104.

Feng, L., Zhang, D., Fan, C., Ma, C., Yang, W., Meng, Y., Wu, W., Guan, S., Jiang, B., Yang, M., et al. (2013). ER stress-mediated apoptosis induced by celastrol in cancer cells and important role of glycogen synthase kinase-3β in the signal network. *Cell Death and Disease* **4**:e715.

Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L., and McCallion, A. S. (2006). Conservation of RET Regulatory Function from Human to Zebrafish Without Sequence Similarity. *Science* **312**:276–279.

Fribley, A. M., Miller, J. R., Brownell, A. L., Garshott, D. M., Zeng, Q., Reist, T. E., Narula, N., Cai, P., Xi, Y., Callaghan, M. U., et al. (2015). Celastrol induces unfolded protein response-dependent cell death in head and neck cancer. *Experimental Cell Research* **330**:412–422.

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., and Engreitz, J. M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**:769–773.

Gardner, L. B. (2008). Hypoxic Inhibition of Nonsense-Mediated RNA Decay Regulates Gene Expression and the Integrated Stress Response. *Molecular and Cellular Biology* **28**:3729–3741.

Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33**:717–728.

Golebiowski, F., Matic, I., Tatham, M. H., Cole, C., Yin, Y., Nakamura, A., Cox, J., Barton, G. J., Mann, M., and Hay, R. T. (2009). System-Wide Changes to SUMO Modifications in Response to Heat Shock. *Sci. Signal.* **2**:ra24–ra24.

Gruss, P., Dhar, R., and Khoury, G. (1981). Simian virus 40 tandem repeated sequences as an element of the early promoter. *Proceedings of the National Academy of Sciences* **78**:943–947.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* **550**:204–213.

Guo, J. L., Yuan, S. X., Wang, X. C., Xu, S. X., and Li, D. D. (1981). Tripterygium wilfordii Hook f in rheumatoid arthritis and ankylosing spondylitis. Preliminary report. *Chinese Medical Journal* **94**:405–412.

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *The American Journal of Human Genetics* **95**:535–552.

Hah, N., Danko, C. G., Core, L., Waterfall, J. J., Siepel, A., Lis, J. T., and Kraus, W. L. (2011). A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells. *Cell* **145**:622–634.

Hah, N., Murakami, S., Nagari, A., Danko, C. G., and Kraus, W. L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. *Genome Research* **23**:1210–1223.

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R., and Eisen, M. B. (2008). Sepsid even-skipped Enhancers Are Functionally Conserved in Drosophila Despite Lack of Sequence Conservation. *PLOS Genetics* **4**:e1000106.

Hay, D., Hughes, J. R., Babbs, C., Davies, J. O. J., Graham, B. J., Hanssen, L. L. P., Kassouf, M. T., Oudelaar, A. M., Sharpe, J. A., Suciu, M. C., et al. (2016). Genetic dissection of the α-globin super-enhancer in vivo. *Nat Genet* .

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**:108–112.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39**:311–318.

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology* **16**:144–154.

Hickey, G., Paten, B., Earl, D., Zerbino, D., and Haussler, D. (2013). HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**:1341–1342.

Hieda, M., Winstanley, H., Maini, P., Iborra, F. J., and Cook, P. R. (2004). Different populations of RNA polymerase II in living mammalian cells. *Chromosome Research* **13**:135–144.

Hieronymus, H., Lamb, J., Ross, K. N., Peng, X. P., Clement, C., Rodina, A., Nieto, M., Du, J., Stegmaier, K., Raj, S. M., et al. (2006). Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10**:321–330.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Res* **34**:D590–D598.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A. (2013). Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155**:934–947.

Hnisz, D., Schuijers, J., Lin, C. Y., Weintraub, A. S., Abraham, B. J., Lee, T. I., Bradner, J. E., and Young, R. A. (2015). Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. *Molecular Cell* **58**:362–370.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* **42**.

Hogues, H., Lavoie, H., Sellam, A., Mangos, M., Roemer, T., Purisima, E., Nantel, A., and Whiteway, M. (2008). Transcription Factor Substitution during the Evolution of Fungal Ribosome Regulation. *Molecular Cell* **29**:552–562.

Holloway, A. K., Bruneau, B. G., Sukonnik, T., Rubenstein, J. L., and Pollard, K. S. (2016). Accelerated Evolution of Enhancer Hotspots in the Mammal Ancestor. *Mol Biol Evol* **33**:1008–1018.

Hughes, J. R., Cheng, J.-F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., Gobbi, M. D., de Jong, P., Rubin, E., and Higgs, D. R. (2005). Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *PNAS* **102**:9830–9835.

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics* **46**:205–212.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**:290–294.

Jonkers, I., Kwak, H., and Lis, J. T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **3**. [Online; accessed 2015-04-13].

Kanazawa, S., Soucek, L., Evan, G., Okamoto, T., and Peterlin, B. M. (2003). C-Myc recruits P-TEFb for transcription, cellular proliferation and apoptosis. *Oncogene* **22**:5707–5711.

Kannaiyan, R., Shanmugam, M. K., and Sethi, G. (2011). Molecular targets of celastrol derived from Thunder of God Vine: Potential role in the treatment of inflammatory disorders and cancer. *Cancer Letters* **303**:9–20.

Katan, Y., Agami, R., and Shaul, Y. (1997). The transcriptional activation and repression domains of RFX1, a context-dependent regulator, can mutually neutralize their activities. *Nucleic Acids Research* **25**:3621–3628.

Khoueiry, P., Girardot, C., Ciglar, L., Peng, P.-C., Gustafson, E. H., Sinha, S., and Furlong, E. E. (2017). Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. *eLife* **6**.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**:182–187.

Kuan, Y.-C., Hashidume, T., Shibata, T., Uchida, K., Shimizu, M., Inoue, J., and Sato, R. (2017). Heat Shock Protein 90 Modulates Lipid Homeostasis by Regulating the Stability and Function of Sterol Regulatory Element-binding Protein (SREBP) and SREBP Cleavage-activating Protein. *Journal of Biological Chemistry* **292**:3016–3028.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**:1812–1819.

Kwak, H., Fuda, N. J., Core, L. J., and Lis, J. T. (2013). Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339**:950–953.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359.

Lässig, M. (2007). From biophysics to evolutionary genetics: Statistical aspects of gene regulation. *BMC Bioinformatics* **8**:1–21.

Lee, B.-K., Bhinge, A. A., and Iyer, V. R. (2011). Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic Acids Research* **39**:3558–3573.

Li, H., Zhang, Y.-y., Huang, X.-Y., Sun, Y.-n., Jia, Y.-f., and Li, D. (2005). Beneficial effect of tripterine on systemic lupus erythematosus induced by active chromatin in BALB/c mice. *European Journal of Pharmacology* **512**:231–237.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., and Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* **352**:600–604.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**:476–482.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Iii, E. J. K., Zody, M. C., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**:803–819.

Lis, J. T., Mason, P., Peng, J., Price, D. H., and Werner, J. (2000). P-TEFb kinase recruitment and function at heat shock loci. *Genes & Development* **14**:792–803.

Liu, J., Lee, J., Salazar Hernandez, M. A., Mazitschek, R., and Ozcan, U. (2015). Treatment of Obesity with Celastrol. *Cell* **161**:999–1011.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**:550.

Ludwig, M. Z., Bergman, C., Patel, N. H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**:564–567.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C. B., Krumm, A., et al. (2017). Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution .

Ma, X., Xu, L., Alberobello, A. T., Gavrilova, O., Bagattin, A., Skarulis, M., Liu, J., Finkel, T., and Mueller, E. (2015). Celastrol Protects against Obesity and Metabolic Dysfunction through Activation of a HSF1-PGC1α Transcriptional Axis. *Cell Metabolism* **22**:695–708.

Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G., and Lis, J. T. (2016). Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Molecular Cell* **62**:63–78.

Marinov, G. K. and Kundaje, A. (2018). ChIP-ping the branches of the tree: Functional genomics and the evolution of eukaryotic gene regulation. *Brief Funct Genomics* .

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J. A., and Churchman, L. S. (2015). Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell* **161**:541–554.

McGaughey, D. M., Vinton, R. M., Huynh, J., Al-Saif, A., Beer, M. A., and McCallion, A. S. (2008). Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. *Genome Res.* **18**:252–260.

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**:495–501.

Mercola, M., Wang, X. F., Olsen, J., and Calame, K. (1983). Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus. *Science* **221**:663–665.

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47**:598–606.

Mikkelsen, T. S., Wakefield, M. J., Aken, B., Amemiya, C. T., Chang, J. L., Duke, S., Garber, M., Gentles, A. J., Goodstadt, L., Heger, A., et al. (2007). Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* **447**:167–177.

Moorthy, S. D., Davidson, S., Shchuka, V. M., Singh, G., Malek-Gilani, N., Langroudi, L., Martchenko, A., So, V., Macpherson, N. N., and Mitchell, J. A. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res.* **27**:246–258.

Mu, T.-W., Ong, D. S. T., Wang, Y.-J., Balch, W. E., Yates III, J. R., Segatori, L., and Kelly, J. W. (2008). Chemical and Biological Approaches Synergize to Ameliorate Protein-Folding Diseases. *Cell* **134**:769–781.

Mullen, K., Ardia, D., Gil, D., Windover, D., and Cline, J. (2011). DEoptim: An R Package for Global Optimization by Differential Evolution. *Journal of Statistical Software, Articles* **40**:1–26.

Nagase, M., Oto, J., Sugiyama, S., Yube, K., Takaishi, Y., and Sakato, N. (2003). Apoptosis induction in HL-60 cells and inhibition of topoisomerase II by triterpene celastrol. *Bioscience, Biotechnology, and Biochemistry* **67**:1883–1887.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* **135**:370–384.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**:83–90.

Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, M. J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N. J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**:526–540.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology* **11**:220.

Ozsolak, F. and Milos, P. M. (2011). RNA sequencing: Advances, challenges and opportunities. *Nature Reviews Genetics* **12**:87–98.

Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K., and Gilad, Y. (2011). A

Genome-Wide Study of DNA Methylation Patterns and Gene Expression Levels in Multiple Human and Chimpanzee Tissues. *PLOS Genetics* **7**:e1001316.

Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* **110**:17921–17926.

Peng, B., Xu, L., Cao, F., Wei, T., Yang, C., Uzan, G., and Zhang, D. (2010). HSP90 inhibitor, celastrol, arrests human monocytic leukemia cell U937 at G0/G1 in thiol-containing agents reversible way. *Molecular Cancer* **9**:79.

Peng, X., Alföldi, J., Gori, K., Eisfeld, A. J., Tyler, S. R., Tisoncik-Go, J., Brawand, D., Law, G. L., Skunca, N., Hatta, M., et al. (2014). The draft genome sequence of the ferret (Mustela putorius furo) facilitates study of human respiratory disease. *Nature Biotechnology* **32**:1250–1255.

Petrovski, S., Gussow, A. B., Wang, Q., Halvorsen, M., Han, Y., Weir, W. H., Allen, A. S., and Goldstein, D. B. (2015). The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLoS Genet* **11**.

Phillips, R. (2015). Napoleon Is in Equilibrium. *Annual Review of Condensed Matter Physics* **6**:85–111.

Pohl, A. and Beato, M. (2014). Bwtool: A tool for bigWig files. *Bioinformatics* page btu056.

Polager, S. and Ginsberg, D. (2003). E2F Mediates Sustained G2 Arrest and Down-regulation of Stathmin and AIM-1 Expression in Response to Genotoxic Stress. *Journal of Biological Chemistry* **278**:1443–1449.

Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., et al. (2006). Forces Shaping the Fastest Evolving Regions in the Human Genome. *PLOS Genetics* **2**:e168.

Pott, S. and Lieb, J. D. (2015). What are super-enhancers? *Nature Genetics* **47**:8–12.

Prescott, S. L., Srinivasan, R., Marchetto, M. C., Grishina, I., Narvaiza, I. n., Selleri, L., Gage, F. H., Swigut, T., and Wysocka, J. (2015). Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163**:68–83.

Qu, J., Hodges, E., Molaro, A., Gagneux, P., Dean, M. D., Hannon, G. J., and Smith, A. D. (2018). Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. *Genome Res.* **28**:145–158.

Quang, D. X., Erdos, M. R., Parker, S. C. J., and Collins, F. S. (2015). Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. *Epigenetics & Chromatin* **8**:23.

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology* **29**:436–442.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* **46**:944–950.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature* **489**:109–113.

Sawai, M., Ishikawa, Y., Ota, A., and Sakurai, H. (2013). The proto-oncogene JUN is a target of the heat shock transcription factor HSF1. *FEBS Journal* **280**:6672–6680.

Scimè, A., Li, L., Ciavarra, G., and Whyte, P. (2008). Cyclin D1/cdk4 can interact with E2F4/DP1 and disrupts its DNA-binding capacity. *J. Cell. Physiol.* **214**:568–581.

Sethi, G., Ahn, K. S., Pandey, M. K., and Aggarwal, B. B. (2007). Celastrol, a novel triterpene, potentiates TNF-induced apoptosis and suppresses invasion of tumor cells by inhibiting NF-κB–regulated gene products and TAK1-mediated NF-κB activation. *Blood* **109**:2727–2735.

Shack, S., Gorospe, M., Fawcett, T. W., Hudgins, W. R., and Holbrook, N. J. (1999). Activation of the cholesterol pathway and Ras maturation in response to stress. *Oncogene* **18**:6021–6028.

Shalgi, R., Hurt, J. A., Krykbaeva, I., Taipale, M., Lindquist, S., and Burge, C. B. (2013). Widespread Regulation of Translation by Elongation Pausing in Heat Shock. *Molecular Cell* **49**:439–452.

Shaulian, E. and Karin, M. (2001). AP-1 in cell proliferation and survival. *Oncogene* **20**:2390–2400.

Shin, H. Y., Willi, M., Yoo, K. H., Zeng, X., Wang, C., Metser, G., and Hennighausen, L. (2016). Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat Genet* **advance online publication**.

Siepel, A. and Haussler, D. (2005). Phylogenetic Hidden Markov Models. In *Statistical Methods in Molecular Evolution*, pages 325–351. Springer-Verlag, New York.

Simpson, K. J., Ranganathan, S., Fisher, J. A., Janssens, P. A., Shaw, D. C., and Nicholas, K. R. (2000). The Gene for a Novel Member of the Whey Acidic Protein Family Encodes Three Four-disulfide Core Domains and Is Asynchronously Expressed during Lactation. *J. Biol. Chem.* **275**:23074–23081.

Smith, E. and Shilatifard, A. (2014). Enhancer biology and enhanceropathies. *Nature Structural & Molecular Biology* **21**:210–219.

Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics* **13**:613–626.

Teves, S. S. and Henikoff, S. (2011). Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. *Genes & Development* **25**:2387–2397.

Todd, D. J., Lee, A.-H., and Glimcher, L. H. (2008). The endoplasmic reticulum stress response in immunity and autoimmunity. *Nature Reviews Immunology* **8**:663–674.

Trott, A., West, J. D., Klaić, L., Westerheide, S. D., Silverman, R. B., Morimoto, R. I., and Morano, K. A. (2008). Activation of Heat Shock and Antioxidant Responses by the Natural Product Celastrol: Transcriptional Signatures of a Thiol-targeted Molecule. *Molecular Biology of the Cell* **19**:1104–1112. 00102 PMID: 18199679.

Urano, F., Wang, X., Bertolotti, A., Zhang, Y., Chung, P., Harding, H. P., and Ron,

D. (2000). Coupling of Stress in the ER to Activation of JNK Protein Kinases by Transmembrane Protein Kinase IRE1. *Science* **287**:664–666.

Vahedi, G., Kanno, Y., Furumoto, Y., Jiang, K., Parker, S. C. J., Erdos, M. R., Davis, S. R., Roychoudhuri, R., Restifo, N. P., Gadina, M., et al. (2015). Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520**:558–562.

Veitia, R. A., Caburet, S., and Birchler, J. A. (2018). Mechanisms of Mendelian dominance. *Clin. Genet.* **93**:419–428.

Vihervaara, A., Mahat, D. B., Guertin, M. J., Chu, T., Danko, C. G., Lis, J. T., and Sistonen, L. (2017). Transcriptional response to stress is pre-wired by promoter and enhancer architecture. *Nature Communications* **8**:255.

Vihervaara, A., Sergelius, C., Vasara, J., Blom, M. A. H., Elsing, A. N., Roos-Mattjus, P., and Sistonen, L. (2013). Transcriptional response to stress in the dynamic chromatin environment of cycling and mitotic cells. *Proceedings of the National Academy of Sciences* **110**:E3388–E3397.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., et al. (2015). Enhancer Evolution across 20 Mammalian Species. *Cell* **160**:554–566.

Villard, J., Peretti, M., Masternak, K., Barras, E., Caretti, G., Mantovani, R., and Reith, W. (2000). A Functionally Essential Domain of RFX5 Mediates Activation of Major Histocompatibility Complex Class II Promoters by Promoting Cooperative Binding between RFX and NF-Y. *Molecular and Cellular Biology* **20**:3364–3376.

Wang, H., Teriete, P., Hu, A., Raveendra-Panickar, D., Pendelton, K., Lazo, J. S., Eiseman, J., Holien, T., Misund, K., Oliynyk, G., et al. (2015a). Direct inhibition of c-Myc-Max heterodimers by celastrol and celastrol-inspired triterpenoids. *Oncotarget* **6**:32380–32395.

Wang, H., Teriete, P., Hu, A., Raveendra-Panickar, D., Pendelton, K., Lazo, J. S., Eiseman, J., Holien, T., Misund, K., Oliynyk, G., et al. (2015b). Direct inhibition of c-Myc-Max heterodimers by celastrol and celastrol-inspired triterpenoids. *Oncotarget* **6**:32380–32395.

Wang, K., Ng, S. K., and McLachlan, G. J. (2012). Clustering of time-course gene expression profiles using normal mixture models with autoregressive random effects. *BMC Bioinformatics* **13**:300.

Wang, X., Sato, R., Brown, M. S., Hua, X., and Goldstein, J. L. (1994). SREBP-1, a membrane-bound transcription factor released by sterol-regulated proteolysis. *Cell* **77**:53–62.

Westerheide, S. D., Bosman, J. D., Mbadugha, B. N. A., Kawahara, T. L. A., Matsumoto, G., Kim, S., Gu, W., Devlin, J. P., Silverman, R. B., and Morimoto, R. I. (2004). Celastrols as Inducers of the Heat Shock Response and Cytoprotection. *Journal of Biological Chemistry* **279**:56053–56060.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**:307–319.

Wilkie, A. O. (1994). The molecular basis of genetic dominance. *J. Med. Genet.* **31**:89–98.

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

Wu, C.-H., Yamaguchi, Y., Benjamin, L. R., Horvat-Gordon, M., Washinsky, J., Enerly, E., Larsson, J., Lambertsson, A., Handa, H., and Gilmour, D. (2003). NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes & Development* **17**:1402–1414.

Xu, X., Wu, Z., Xu, C., Ren, Y., and Ge, Y. (2003). Observation on serum anti-double stranded DNA antibodies of tripterine in systemic lupus erythematosus of (NZBxW)F1 mice. *Annals of the Rheumatic Diseases* **62**:377–378.

Xu, Y., Sengupta, P. K., Seto, E., and Smith, B. D. (2006). Regulatory Factor for X-box Family Proteins Differentially Interact with Histone Deacetylases to Repress Collagen *a*2(I) Gene (COL1A2) Expression. *Journal of Biological Chemistry* **281**:9260–9270.

Yan, G., Zhang, G., Fang, X., Zhang, Y., Li, C., Ling, F., Cooper, D. N., Li, Q., Li, Y., van Gool, A. J., et al. (2011). Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nature Biotechnology* **29**:1019–1023.

Yang, H., Chen, D., Cui, Q. C., Yuan, X., and Dou, Q. P. (2006). Celastrol, a Triterpene Extracted from the Chinese "Thunder of God Vine," Is a Potent Proteasome Inhibitor and Suppresses Human Prostate Cancer Growth in Nude Mice. *Cancer Res* **66**:4758–4765.

Yang, S., Oksenberg, N., Takayama, S., Heo, S.-J., Poliakov, A., Ahituv, N., Dubchak, I., and Boffelli, D. (2015). Functionally conserved enhancers with divergent sequences in distant vertebrates. *BMC Genomics* **16**:882.

Yang, Y., Gu, Q., Zhang, Y., Sasaki, T., Crivello, J., O'Neill, R. J., Gilbert, D. M., and Ma, J. (2018). Continuous-trait probabilistic model for comparing multi-species functional genomic data. *bioRxiv* page 283093.

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**:556–559.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R. A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nature Genetics* **39**:1512–1516.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., et al. (2018). Ensembl 2018. *Nucleic Acids Res* **46**:D754–D761.

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., and Flicek, P. R. (2015). The Ensembl Regulatory Build. *Genome Biology* **16**:56.

Zhang, T., Li, Y., Yu, Y., Zou, P., Jiang, Y., and Sun, D. (2009). Characterization of Celastrol to Inhibit Hsp90 and Cdc37 Interaction. *Journal of Biological Chemistry* **284**:35381–35389.

Zhang, Y., Geng, C., Liu, X., Li, M., Gao, M., Liu, X., Fang, F., and Chang, Y. (2016). Celastrol ameliorates liver metabolic damage caused by a high-fat diet through Sirt1. *Molecular Metabolism* **6**:138–147.