Title
Computational and Experimental Investigations of the Principles of Eukaryotic Transcriptional Regulation Before, During, and After Open Complex Formation

Author
DAVIS, MATTHEW D.

Publication Date
2012

Peer reviewed|Thesis/dissertation

Computational and Experimental Investigations of the Principles of Eukaryotic Transcriptional Regulation Before, During, and After Open Complex Formation

by

Matthew D. Davis

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Eisen, Chair

Professor Rachel Brem

Professor Nipam Patel

Professor Sandrine Dudoit

Fall 2012

Abstract

Computational and Experimental Investigations of the Principles of Eukaryotic Transcriptional Regulation Before, During, and After Open Complex Formation

By

Matthew D. Davis

Doctor of Philosophy in Molecular and Cell Biology

With Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Michael Eisen, Chair

In this work, I describe my doctoral work studying the regulation of transcription with both computational and experimental methods before, during, and after recruitment of RNA Polymerase II and the formation of the open complex at gene promoters. This works spans investigation of the combinatorial binding of sequence specific distal enhancer binding proteins in the developing fly embryo the role of the yeast AAA-ATPase YTA7 in reorganizing the nucleosome positioning at the highly transcribed loci, and the role of the AFF scaffold-associated members of the human super elongation complex.

This work is dedicated to the relationships with friends, family, and other dear persons, in sickness and in health, that have weathered my selfish neglect in pursuit of these studies, but especially to the ones that did not.

I would like to acknowledge the following individuals for providing a collection of guidance, helpful hands, and fruitful discussions in recent years:

Tom Alber, Eileen Bell, Mark Biggin, Rachel Brem, Colin Brown, Seemay Chou, Charles Denby, Sandrine Dudoit, Mike Eisen, Sarah Ewald, Malik Francis, James Fraser, Aaron Hardin, Emily Hare, Melissa Harrison, Tommy Kaplan, Terry Lang, Xiao-Yong Li, Laura Lombardi, Susan Lott, Rich Lusk, Edward Marcotte, Andy Mehle, Nipam Patel, Brant Peterson, Dan Richter, Matt Taliaferro, Jacqueline Villalta, Oh-Kyu Yoon, and many other members of the Eisen Lab, the Department of Molecular and Cell Biology, and the community of the University of California.

"In the last five years we are learning to do molecular genetics directly by sequencing the DNA. Sanger, and these people, who show how you actually sequence the DNA and find the base change. Whereas the attempts to use genetics could be interpreted as just cheap ways of trying to sequence the DNA."

Sydney Brenner

February, 1976[1]

Introduction

Transcriptional Regulators Decipher the Animal Body Plan Encoded in Genomic DNA

The blueprints for all metabolic products of living cells, including nucleic acid, protein, fats, sugars, are all ultimately encoded in DNA. The genomes of living organisms encode not just the information for the construction of these metabolic products, but also the information for the timing and ordering of production, organization, and decay of these molecules that underlie the differentiation of cell types and development of organismal body plans. The flow of information in living systems thus is from DNA to metabolic intermediates such as messenger RNA to building blocks of cells and tissues to fully developed whole organisms. It is a central scientific challenge to understand how this information is organized to constitute living beings. The elucidation of the genetic code made clear how the information was immediately organized[2-4], but higher-order organization of the information for gene expression, molecular half-life, and post-translational modification is not directly encoded. Turing noted that through simple diffusion-reaction mechanisms, a process of cell fate restriction encoded in the genetic body plan of animals could be parsed throughout development[5]. With the development of molecular biology and eventually the direct sequencing of DNA, great inroads were made in understanding this flow of information, but not until the end of the 1990s were scientists first able to characterize and analyze the content of an entire animal genome[6].

The era of eukaryotic genome sequences began with the completion of the yeast genome sequence by an international consortium in 1996[7]. Roughly ten years later, the list of available eukaryotic genome sequences had expanded to include the model organisms of worm, fly, and mouse[6,8,9], as well as the human genome[10]. Surprisingly, the analysis of the human genome found far fewer genes encoded than was expected based on estimates extrapolated from other organisms. For example, the genome of *C. elegans* contains roughly as many genes as the human genome, though the two organisms differ by roughly nine orders of magnitude in cell number per individual. The question of the origin of metazoan organismal complexity cannot be answered simply by content of the genes in the genome, but perhaps it can be explained by the combinations in which those genes are used to produce different cell types, tissues, and organs over the course of the developmental plan of an organism[11-13]. The yeast genome harbors approximately 6,000 open reading frames comprising 70% of the nucleotide sequence of the genome, but the coding sequence of the human genome comprises only 1.2% of the complete nucleotide sequence[14],

consistent with the notion that non-coding sequence underlies organismal complexity. This also suggests that the complexity of the organism scales not with genome size, but with ability of the genome to generate combinations of gene expression that result in diverse cellular outcomes.

Following the completion of closely related genome sequences[9,15-19], early estimates of the proportion of regulatory elements in model genomes were made by computationally assessing conservation of non-coding elements across related genomes. Comparative genomics led the way in identifying regulatory elements on a genomic scale, though these methods did little to reveal the mechanisms of regulation[20,21]. Even if conserved regulatory regions are responsible for the derivation of a cell type or tissue layer, sequence conservation methods cannot readily determine at which stage of development the region is active nor identify the target genes at this stage. In cases where population genetic samples are available and sufficient variance exists in the activity of the regulatory elements, association studies can demonstrate an association with phenotypes and their underlying genotypes regulated by these elements[22-30]. Again, the question of which genes and which developmental stage are key is not readily answered. However, these methods are complemented by genomic molecular assays for the discovery of functional elements including ChIP-chip[31,32], ChIP-Seq[33], DNaseI Hypersenstivity[34], and Hi-C[35,36] which can be used to discover the stage of activity, and in many cases the target genes of interest.

Mechanisms of Transcriptional Regulation

The first advances in understanding the regulation of gene expression came with Jacob and Monod's dissection of the lac operon by induction and inhibition[37], and over fifty years later, transcriptional regulation remains a field of active discovery. Yeast genetics identified many transcriptional regulators of metabolism including the cell cycle[38] and reproductive pathways[39]. Developmental genetics linked the role of homeotic factors and morphogens to transcription[40-44], which facilitated our understanding of combinatorial regulation of gene expression in animals by distal transcriptional enhancers[45]. Biochemists purified and characterized the DNA binding domains and identified the components of the RNA Polymerase II basal transcription machinery[46]. More recently, genomic studies of RNA Pol II localization have revealed pervasive pausing as a critical step in gene expression[47-50], and characterizations of the regulation of elongation, persistence, and transcriptional rate have all become possible with the advent of high-throughput sequencing methods[51,52]. Gene regulation also takes place at the level of RNA stability[53-56], translational efficiency[57-59], and the post-translational modifications of proteins and their stability, but transcriptional regulation is theoretically and empirically the principally important domain of gene regulation[60,61]. The modes of transcriptional regulation fall into three sets of complementary factors and processes: 1) chromatin accessibility and remodeling, 2) distal enhancer binding factors and the recruitment of the basal RNA Polymerase II transcriptional machinery to form the open complex, and 3) RNA Polymerase II elongation and persistence.

Genomic characterization of histone modifications and nucleosome positioning revealed enrichment of histone variants at the sites of active transcription[62]. Most notable is the presence of the histone H2A variant H2A.[63]Z, which is often enriched at the +1 or -1 nucleosome position relative to the transcription start site (TSS). It is thought that this non-canonical histone variant is less tightly associated with the DNA itself, and is more easily evicted by histone remodeling proteins. One such mechanism is the recruitment of a histone acetyltransferase (HAT)[64]. For example, pCAF, a human HAT, can be recruited by the TATA-binding protein (TBP) in complex with either TATA or SAGA box proteins[65]. Another is the use of ATP hydrolyzing enzymes such as the SWI/SNF complex[66]. These proteins use the energy from ATP hydrolysis to slide nucleosomes and ensure proper spacing, but can also act to expel nucleosomes, facilitating access by the basal transcriptional machinery. In addition to the role of chromatin in facilitating or encumbering access to the DNA by transcription factors and RNA Polymerase II, chromatin modifications mark different types of regions within coding regions as well[62]. During the course of active transcription, nucleosomes remain associated with the transcribed DNA, and must handled by the transcriptional machinery. The mechanisms of these interactions are poorly understood, but one role is for AAA-ATPases to transiently dissociate the nucleosomes from the DNA, allowing ready passage of the polymerase complex through the locus.

While existing chromatin state influences transcriptional initiation and efficacy, the existence of any sequence-specific signal directing the positioning of nucleosome[67,68] in the genome has been controversial[69-71]. There appear to be sequences that favor the bending of DNA required for nucleosome wrapping, but whether this signal is sufficient to position the nucleosomes in the chromatin amid the much stronger sequence signals encoding the recognition sequences for DNA binding domains of transcription factors is unclear. What is clear is that sequence specific DNA binding domains are capable of interpreting the probabilistic code of regulatory sequence in the genome.

Both promoter proximal and distal enhancer binding transcription factors rely on sequence specific binding domains to recognize their binding sites.[72-75] Depending on the binding domain, these sequences are typically less than a dozen base pairs long, although oligomerization of cooperating factors can increase the length of the binding site[76-78]. These sites are not deterministically bound, but rather probabilistically and transiently occupied by their binding domains. *In vitro* assays of binding kinetics suggest that DNA binding domains bind their ligands in a cooperative manner in accordance with Hill kinetics, with the Hill coefficient varying for each pair of DNA binding domain and ligand[79-82]. The probability of binding is thus a function of both the affinity of the sequence for the DNA binding domain and the concentration of the factor present *in vitro*. The primary sequence affinity of DNA binding domains can be determined *in vitro* by selection assays, as well by other methods described below. The compendium of the selected bound sequences can be convolved into a probabilistic weight matrix (PWM)[83-85], which can be shown to have strong power to predict the probabilistic binding kinetics of factors both

empirically and by principles of information theory. These values, however, are most applicable *in vitro*, as many other forces influence binding affinity *in vivo*.

The concentration of the factor *in vivo* has often been presumed to be in radical excess of the concentration of DNA binding domains, suggesting that when a transcription factor is induced, its sites are generally occupied[86]. Reason to doubt this assumption issues from the observation that transcription factor concentration varies from tissue to tissue in a continuous fashion, and from recent work studying the kinetics of *in vivo* binding of glucocorticoid receptor[87-89]. The central focus of the third chapter of this work is an *in vivo* titration experiment that aims to inform the relationship between the continuum of transcription factor concentrations in the fly embryo and binding of transcription factors to developmental distal enhancer regions.

In addition to chromatin status, primary sequence affinity, and factor concentration, the availability of transcriptional cofactors can also influence the binding of a transcription factor to its recognition sequence. In many cases, such as the dimeric leucine zipper family, the transcription factor must oligomerize in order to recognize its sequence ligand. In some cases, this is heterodimerization with other DNA binding proteins, and in other cases, the factor can facultatively oligomerize with cofactors that do not binding DNA. In some cases, such as has been studied in atomic detail with hox proteins and their cofactors[90-94], the binding recognition or biophysical mechanism of binding is altered. In yet other cases, the cofator is required for stabilization or association with a larger complex of transcriptional proteins[95,96].

The mechanisms of transcriptional activation are complex, but generally understood. Upon binding of an activator protein to a distal enhancer region, provided there is not sufficient repressive activity present, the activator will either directly or via a co-factor, bind to the basal transcriptional machinery, often the mediator complex, which is in turn competent to recruit and stabilize the basal sequence specific transcription factors such as TFIID and RNA Polymerase II[97]. Mechanisms of repression can be roughly characterized as any molecular process that disrupts the process of activation. This can involve direct competition for recognition sequences shared by activators and their antagonistic repressors[98-100]. Repressors may also form DNA loops that exclude the region bound by an activator, sterically interfering with the activation of the basal complex. Repression may also occur in the presence of excess activator by a mechanism called “squelching” where the excess activator sequesters necessary cofactors and outcompetes the activator molecules positioned at the proper regulatory regions[81]. Many other models for repression should be considered plausible, though the evidence the relative efficacy of each model of repression is scarce. In sum, repression is interference with the course of activation.

The phenomenon of paused polymerase at the *Drosophila* HSP70 gene was first described in 1992[101], and genomic localization data for RNA Polymerase II has since shown pervasive RNA Polymerase II pausing in several models[102]. The function of

paused polymerase is still a matter for study, but pausing is relieved in the course of transcriptional elongation. The phosphorylation of the Ser2 residues of repeats in the C-terminal tail of RNA Polymerase II is necessary for elongation[103]. This phosphorylation is catalyzed by the CDK9 kinase[104]. The context of this regulatory step is still an active pursuit of research, but recent work suggests that the CDK9/P-Tefb complex is recruited by the bromo-domain protein Brd4[105,106]. These proteins are scaffolded at many loci by one of the AFF scaffolds of the Super Elongation Complex (SEC), which are also known to scaffold a number of other factors that are positively associated with active and efficient transcript elongation[107-110]. Other well-studied factors such as the transcription factor c-Myc have also been associated with efficient phosphorylation of RNA Polymerase II, and many of the genes in this process were first identified for their roles in tumorigenesis, reflective of their ability to broadly influence transcription of many genes.

These intertwined molecular processes together constitute the pre-initiation, stabilization, and elongation phases of eukaryotic transcriptional regulation. The recognition of gene expression by combinatorial binding to distal enhancer sequences is a pervasive feature of metazoan gene expression, as it is the basis for tissue differentiation during development. The canon of comparative work since the discovery of the homeotic genes also suggests that the evolution of gene expression acts largely through selection on the cis-regulatory logic of distal enhancer sequences[11,29], though many cases of evolution in the coding sequences of proteins active in interpreting the regulation of genes have been illustrated[111-113]. More recently, there is evidence accumulating for cell type specific cis-regulation at the basal promoter[114] and in the transcriptional elongation complex, and these processes may yet prove powerful to effect cellular differentiation, especially early in development.

Methods

This section describes and reviews the basic assumptions of the methods used to generate and analyze data relevant to the research described in subsequent chapters.

Biochemical and Molecular Methods

The inference of binding sites recognized by sequence specific transcription factors has greatly facilitated our understanding of higher order organization of regulatory information in eukaryotes. Binding motifs can be inferred from both *in vitro* and *in vivo* data. DNA footprinting is a protection assay where a pool of sequence, typically amplified from a genomic region of interest, is incubated with a transcription factor and the DNA is digested[115]. The protected fragments are detected with a gel assay, and individual protected sequences can be inferred. This is similar in principle to the DNA gel-shift, which is used to illustrate binding and estimate the affinity of particular DNA ligands for their binding domains. These two methods proved very powerful and convolutions of the sequences inferred from footprinting have been used to build PWMs with high predictive power. However, these methods are

relatively low-throughput, and are limited by the pool of sequences included in the binding pool. The SELEX assay is a selective enrichment binding assay that enriches high-affinity sequences by iterative selection *in vitro*[116-118]. This method has the advantage of inputting large, unbiased sequence pools for selection, and thus can infer more complex and accurate PWMs. These methods typically assay the binding of only one protein, though in theory mixtures of protein or extract could be used. Bacterial one-hybrid [119]assays and protein-binding microarrays[120] also provide high-throughput methods for binding site enrichment, but none of the data in the work described here were derived from these methods. Finally, ChIP-chip[31,32,121] and ChIP-Seq are *in vivo* methods for binding site identification. The collection of bound regions in a ChIP sample, especially when conditioned on the binding to or near to known functional regions, can allow inference of the actual *in vivo* binding sites and PWMs can be constructed from the convolution of these sequences.

Chromatin immunopurification (ChIP) requires the generation of a factor-specific primary antibody. The spurious cross-recognition of paralogous binding domains is a concern, and to this end multiple antibodies can be generated to regions without high conservation and their results compared. The antibodies used in Chapter 3, for example, were generated with constructs previously validated to give good correspondence between immunopurification of both N-terminal and C-terminal segments of the studied factors[122]. Even with cross-validated antibodies, an orthogonal dataset such as binding of a cofactor or presence of the known binding sites in the ChIP regions can be reassuring, as ChIP assays have a high false-positive rate. Statistically, this can be controlled by establishing a false discovery rate for the dataset and considering each data point accordingly[123]. For reasons that are well understood, open chromatin regions, typically near highly expressed genes, are commonly enriched in likely false positive regions. Whether or not the binding of factors to these regions is genuine or an artifact of crosslinking or purification is difficult to discern, but some effort can be taken to consider a negative control sample. For example, in the case of tagged proteins, a no-tag ChIP experiment can be performed and used as a background dataset for the determining enrichment. This approach was taken in with the yeast ChIP samples in Chapter 4. Another form of control is to ChIP a sequence-specific factor with orthogonal function, an approach employed in Chapter 4 by comparing the binding of the chromatin insulator CTCF as a control for the members of the SEC.

Data generated by high-throughput sequencing methods were used throughout this work. In each case, the data were generated with the Illumina short-read sequencing platform. The ChIP-sequencing protocol developed for the work in Chapter 3 required extensive optimization of the collection and isolation of chromatin, and due to the low input concentrations, the standard Illumina protocol was also optimized. In some cases, carrier sample was used to facilitate proportionally high recovery. Generally, ChIP-sequencing samples of human chromatin required over 20 million reads for reliable data, whereas fly samples generated good results with as few as 2.8 million reads, and yeast samples required far less. For RNA sequencing used to profile gene expression of human cell culture and single fly embryos, read

depth continued to improve signal and resolution of the data past 30 million reads per sample.

Genetics and Developmental Methods

The *Kruppel* mutant flies generated for use in Chapter 3 took advantage of the phi-C31 integrase system[124], which precisely inserts transgenic fragments into engineered landing sites. This avoids inconsistent expression from position effect, and according to RNA-seq profiling of the mutant line, provided nearly precise over-expression of the inserted locus.

In situ hybridization and protein antibody staining were used in Chapters 2 and 3 to assess the localization of both mRNA and transcription factor proteins. These methods have the advantage of providing cellular resolution in the localization and quantification of the gene products measured. However, the quantification is relative within the samples. In the case of the stains collected in this work, as opposed to the data generated by the Berkeley Drosophila Transcription Network Project (BDTNP) and analyzed here, mixed samples were pooled and stained together, using the apparent molecular phenotypes of reference patterns to identify the sample variants after staining. In this manner, experimental variation was minimized and the relative values in each line given high confidence.

Computational Methods

The methods for binding site identification and calculation of primary sequence affinity used here are not novel in theory nor practice[83,125]. The identification of binding sites used the highest information PWM available for the factor of interest. For example, for identifying BCD sites and calculating regional sequence affinity, a high-quality PWM derived from DNA footprinting data was used[126], but for KR, a higher-information matrix derived from SELEX data[122] was preferred. More sophisticated models employ a thermodynamic model of competition or cooperation of other chromatin proteins or states, such as the binding sites of known cofactors or profiles of chromatin accessibility[127,128], but this work employs a simpler additive model of sequence affinity.

In Chapter 2, I used Classification and Regression Trees (CART)[129] to partition the embryo into regions that are similar with respect the expression of over 100 factors, while integrating genomic binding profiles for relevant transcription factors as priors. Other such efforts to predict gene expression logic from expression and binding data have been taken one of two different approaches. Dynamical systems of coupled equations have been constructed, simulated, and optimized chiefly by John Reinitz and colleagues[130,131]. These models have the advantage of explicitly addressing the known phenomena of non-linearities such as feed-back and feed-forward loops. However, these models inherently explore a vast landscape of possible regulatory relationships, and must be limited by assumptions of biological architecture in an effort to avoid entrapment in local minima of biological irrelevance. A second class of models integrates sequence affinity, binding, and

coexpression into a single thermodynamic model, and the parameters of which can be trained against observed data. These models have yielded biologically relevant insights, such as the predictive power of DNase I accessibility as a marker of open chromatin[127] and support for short-range repression as the predominant mode of repression in fly segmentation network[128]. However, these models too assume biological architecture, for example which factors are capable of activation or repression, which I aimed to infer.

Summary

In the following chapters, I describe my work investigating the regulation of transcription by combinatorial binding of sequence specific distal enhancer binding proteins in the developing fly embryo, as well as the role of the yeast AAA-ATPase YTA7 in reorganizing the nucleosome positioning at the highly transcribed loci at which it facilitates expression, and the role of the AFF scaffold-associated members of the human super elongation complex (SEC) in facilitating transcript elongation after the phosphoactivation of the C-terminal tail of RNA Polymerase II.

In Chapter 2, I infer the roles of maternal and gap factors in the expression of *Drosophila* patterning genes using CART to integrate genomic binding and spatially-resolved co-expression data in the embryo. I show evidence that gap patterns are readily partitioned and logic can be inferred that is consistent with the complex roles of morphogen concentration, cross-repression, temporal dynamics, and context-dependent roles of factors as activator and repressor. However, due in part to the limitations of my model to describe autoregulation and to the significant autocorrelation in the interdigitated patterns of the pair-rule genes, these decision tree models fail to make compelling predictions of pair-rule regulatory logic. I offer in conclusion from this work that in addition to pervasive autoregulation in the segmentation network, a finer granularity in the definition of cis-regulatory element would benefit the development of future models.

In Chapter 3, I address the role of transcription factor concentration *in vivo* by assaying genomic binding and expression levels alongside the patterning of canonical targets in two dosage series of the transcription factors *bicoid* and *Kruppel*. I show that there is pervasive sensitivity to dosage of these factors at the level of genomic binding and expression of nearby genes. Further, dosage sensitivity is underwritten by primary sequence affinity, and suggests a model of *in vivo* transcription where factor concentration is not in radical excess of binding sites. However, the model suggests that many functional sites are effectively saturated, perhaps due to forces beyond primary sequence affinity.

Finally, in Chapter 4, I briefly describe vignettes of my work in two collaborations. By integrating genomic binding, expression, and MNase sensitivity data, my collaborators and I have established a wide role for the AAA-ATPase YTA7 in regulating nucleosome spacing and positioning at highly expressed loci in budding yeast. This regulation at the level of chromatin remodeling and expulsion of histone facilitates active transcription after the recruitment of RNA Polymerase II to the

promoter of the gene. And in my second collaboration, I have integrated genomic localization data for the members of the AFF scaffold-associated SEC with RNA-seq data comparing conditions of wild-type and knock-down expression of these factors in human HeLa cells. This work studying the regulation of transcription during and after the elongation of RNA Polymerase II suggests that another level of combinatorial regulation exists downstream of open complex formation, and that some, but not all, of the members of the SEC may travel past the first exon of highly transcribed loci to facilitate efficient transcript elongation.

In summary, I provide a study of transcriptional regulation before, during, and after recruitment of RNA Polymerase II and the formation of the open promoter complex. This ranges a contemporary survey of transcriptional biology that leaves me with several conclusions. It is clear that appropriate models for combinatorial regulation at distal enhancers need to allow for multivariate cooperativity of transcription factors themselves and the genomic context in which they operate. These models should be as open as possible, placing constraint only where there is strong evidence to do so. For example, factor concentration should not be assumed to have a binary nor monotonic role in the activity of a factor. In addition to the complexities of RNA Polymerase II recruitment and open complex formation, the genomic context of the transcribed locus, along with the proteins that guide and regulate the processes of elongation and termination, should also be considered in these models. Contemporary methods such as global run-on sequencing (GRO-Seq)[132] and nascent transcript sequencing (NET-Seq)[51,52] provide additional tools with which to study these mechanisms, but it is worth mentioning that the most of discoveries in each chapter of this work have relied on modern sequencing technology leveraged to gain insight into the complex, abstract, and unseeable mechanisms of the cell. In recognition of this, I began this dissertation with a quotation from Sydney Brenner regarding the inference of the genetic code and role of direct sequencing in the prospects for future discoveries in molecular biology.

Chapter 2

Inference of Combinatorial Logic of Transcriptional Regulation in the *Drosophila* Blastoderm with Classification Trees

Introduction

A foundational discovery of early molecular biology came with the discovery of the lac operon and the induction of gene expression[37]. Though the experimental system that led the way in our nascent understanding of gene regulation required passage of genetic material from one individual to another, it has since been established that eukaryotic genes are coordinately expressed by a compendium of trans-acting transcription factors. As our understanding of the organization of genes and regulatory elements in genomes has developed through the era of molecular biology and genomics, it has become clear that the information encoded in the primary sequence is read by the state of this trans-acting network[12]. Unlike the fundamental genetic code, the regulatory code does not have a simple grammar, but instead integrates probabilistic regulatory logic contained in regions positioned in *cis* to the loci under control.

By encoding short sequences recognized by the DNA binding domains of transcription factors, cis-regulatory modules (CRMs) direct the recruitment of activating and repressing domains, which in turn associate with basal transcription machinery to determine the transcriptional state of the target locus. In yeast, this regulatory logic is encoded in short promoter regions near the transcription start site of the gene[73,133], but in metazoans, distal CRMs direct expression of the locus, often from distances greater than the length of the gene itself and with much more sophisticated mechanics[97]. This complexity increases the field of possible biophysical mechanisms for gene regulation. However, only a few examples of such regulation are well-documented[134-137], and direct observation of these mechanisms remains an active pursuit[138]. Whereas the regulatory logic in microbes like yeast can be predicted by inferring the roles of trans-acting factors as activators or repressors[139], the more complex architecture of metazoan CRMs presents a challenge for inference and prediction. Additionally, though yeast exhibits different trans-acting network properties throughout the cell cycle, between the mating types, and in varying media environments, metazoans face a fundamentally more complex task in combinatorial gene regulation. From a set of isogenic pluripotent cells, metazoans must derive various cell and tissue types, precisely arranged in space and time to form coordinated organs and limbs that function as one body. Thus, heterogenous isogenic microbial populations differ fundamentally from metazoan cells as they lack the non-autonomous influence metazoan cells assert on one another in the same body.

The blastoderm of the fruit fly *Drosophila melanogaster* is the most established model system for understanding the transcriptional regulatory circuits underlying tissue differentiation in a metazoan. Early work characterized the factors of the trans-acting network with classical genetics, biochemistry, and molecular biology[42,44,135,140-143]. Reporter constructs identified activating sequences required to generate a given expression pattern, and consensus binding sites could be inferred with DNA footprinting. As measurement accuracy improved, it became

clear that there were quantitative continua of transcription factor concentrations distributed throughout the nuclei of the blastoderm[86,144]. As genome sequences became available, comparative genomic and genome-wide molecular assays brought a more complete understanding of the architecture of CRMs and their target genes[18,19,122,145-147]. Thus, the earlier qualitative assays that facilitated our fundamental understanding of metazoan transcriptional regulation gave way to computational and genomic methods to study the complexity of development.

The initial conditions of this system are specified by the deposit of maternal factors into the egg prior to fertilization, and the system of regulators then plays forward through development. Typically, the system has been reduced to study by axis of development, either anterior-posterior or dorsal-ventral, or alternatively by grouping the factors according to their time of onset as maternal, gap, pair-rule, or segmentation factors[148-151]. Here, I integrate datasets produced by the Berkeley Drosophila Transcription Network Project (BDTNP)[122,146,152-154], generated with the goal of considering the full complexity of the blastoderm transcriptional network.

The BDTNP gene expression atlas contains quantitative expression data for over 100 blastoderm factors, including a nearly complete set of the transcription factors expressed at this stage in development. The atlas spans seven intermediate time points throughout the 140 minutes prior to gastrulation, capturing the quantitative variation in the concentration of maternal, gap, pair-rule, and segmentation factors (Figure 1). The data are collected in three physical dimensions with a two-photon confocal microscope, and registered into the atlas of 6,078 nuclei containing both protein and mRNA localization data[152-154]. The registration of thousands of embryos into a single atlas allows for relative quantitative comparison across the nuclei and between factors and time points and thus is the only dataset of its kind.

In addition to the gene expression atlas, the BDTNP also generated genomic binding data for 21 transcription factors in the blastoderm. These 21 include maternal, gap, pair-rule, and ubiquitous factors, as well as representing AP and DV axis determinants. These data, especially in combination with the genome sequences of 12 *Drosophila* species generated by the Berkeley Drosophila Genome Project[19], provide information about which factors are able to regulate which target loci in the genome. However, owing to both the false positive rate inherent to chromatin IP data and to the prevalence of real biophysical, but functionally irrelevant opportunistic binding of transcription factors[122], these data alone are insufficient to allow insightful modeling of the regulatory network.

Here, I leverage the substantial collection of data generated by the BDTNP to infer regulatory relationships in the *Drosophila* blastoderm. Using the rich matrix of mRNA and protein localization data from the gene expression atlas, while considering the genome-wide ChIP-chip binding data, I show that the blastoderm can be readily partitioned into contiguous segments that share a regulatory architecture with respect to the expression of a given target gene. The early network of maternal and gap gene expression is well-described by this method, recapturing existing regulatory knowledge. Additionally, I was able to predict previously

undescribed time-dynamic regulatory relationships between maternal, gap, and pair-rule gene products, generating specific and testable hypotheses about gene regulation. While the algorithm developed is also able to describe portions of the later pair-rule gene expression patterns, these patterns present challenges to the method which themselves call into question our theoretical conception of CRM structure and function.

Results

Classification Trees Accurately Describe Known Regulatory Relationships of Maternal and Gap Genes

The organizational task solved by the developmental plan encoded in the genome of the fly is to differentiate similar cell types in contiguous sections of the embryo prior to gastrulation. If the fundamental cell fates have not been specified by the beginning of gastrulation, then the layers of endoderm, mesoderm, and ectoderm cannot be properly specified. To the extent that a given gene product determines these fates, then cells similar in the concentration of that factor are similar to each other. Or as formulated here, each cell contains a vector of gene product concentrations, and tissue similarity is determined by the overall similarity of these vectors. However, some factors are more determinative than others given the context of their co-expression with other factors. The task is then to understand how each gene product is distributed across the nuclei of the embryos cells as a function of that gene's regulators. In this view, we can conditionally dissect the embryo into groups of cells that are similar with respect to the concentration of a given factor, and then ask how the concentrations of given regulators covary, if at all. To this end, each factor concentration vector in the gene expression atlas was partitioned with a classification tree[129] into similar groups of cells (see methods for details) and the rule sets that determined the branch points of the tree were interpreted as regulatory logic.

For each classification tree, the gene product of interest was predicted by other factors active before and/or during the expression of the target gene. For example, gap genes were predicted by the concentration vectors of maternal and other gap gene products, whereas pair-rule target genes were predicted by maternal, gap, and pair-rule concentration vectors. Additionally, the prediction vectors were excluded from a given classification tree if the BDTNP ChIP data suggested that there was no binding near the target gene.

The classification tree method uses a linear model to predict the target vector, and the divides the observations in the prediction vector above and below the mean of the target value. In this way, the data are partitioned iteratively until there is the refinement of the residual in the linear model no longer improves, at which point the iterations cease and the model produces a "leaf" partition of the classification tree. If there is no structured information in the data matrix, then no branches can be made and a null tree results. Trees with a larger number of leaves represent a more complex conditional data space.

In classifying the nuclei of the gene expression atlas by a given pattern, the algorithm did not produce a null tree for any target gene pattern. Rather, the partitions of the trees fall overwhelmingly into spatially contiguous patterns of nuclei (e.g. Figures 2-5), though there is no explicit spatial information being considered by the algorithm. Thus, the partitioning algorithm can successfully resolve spatial segments of the embryo with respect to a given gene product. Further, the algorithm is sensitive to quantitative variation in both the predictor concentration matrix and the target gene concentration vector. For example, the early expression pattern of the gap gene *hunchback* is driven directly by the Bicoid (BCD) morphogen gradient in a concentration-dependent manner[155]. At the anterior of the embryo, high levels of BCD induce high *hunchback* expression, in the trunk of the embryo, intermediate levels of BCD cooperate with *hunchback* autoregulation to create a sharp expression border, and in the posterior of the embryo, low levels of BCD are insufficient to induce *hunchback*. This morphogen pattern has been referred to as the "French Flag" model by Wolpert[156]. It is accurately recaptured by the partitioning of the embryo's *hunchback* pattern by three separate levels of BCD, despite the presence of dozens of other maternal and gap factor concentration vectors in the predictor matrix (Figure 2).

The classification trees are subject to the limitation of the data they partition, as exhibited by the tree describing the regulatory logic for *giant* expression in the early blastoderm (Figure 3). In this case, the tree correctly identifies Kruppel protein (KR) as the principle repressor of *giant* in the trunk of the embryo and BCD as the principle activator of *giant* in the anterior of the embryo[157]. However, the tree also falsely attributes a repressor role to BCD in the anterior terminus of the embryo. This owes to the fact that the expression atlas does not contain information for the gene products of the *torso* locus, which are known to post-translationally modify BCD in a way that inhibits its role as an activator, and Torso protein is localized to the anterior terminus[158]. Thus, what is actually a lack of BCD activity in the terminus is attributed to very high levels of BCD, falsely suggesting a repressor role. Similarly, the atlas data for the posterior regulatory Caudal (CAD) is relatively poor at this time stage. Since the BCD and CAD gradients oppose one another, the lack of information in the CAD data is supplanted by the BCD data as a proxy.

Despite the imperfections of the dataset, however, the classification trees are able to parse gene expression of both high and low complexity. In the case of the gap gene *knirps*, the classification tree accurately describes the regulation of the locus in both the anterior-posterior and dorsal-ventral system. The activation of *knirps* by BCD, Hunchback (HB), and Tailless (TLL) proteins in the anterior is inhibited by the Snail repressor (SNA) in the anterior dorsal-ventral plane[159]. In the posterior of the embryo, *knirps* regulation is described by repression by GT and activation by TLL. However, the principle activator of *knirps* in the posterior domain is misattributed to a negative relationship with the BCD gradient, again reflecting the weak CAD data in the expression atlas.

A Novel Regulatory Relationship of the hunchback Locus

In the progression of the blastoderm stage, increased pattern complexity becomes common. The hunchback pattern that begins as a single domain expressed as a product of the BCD gradient and cooperative autoregulation resolves into three clear expression domains by the mid-blastoderm stage. As this complexity develops, the pair-rule genes become active regulators and the responsibilities of regulation of each locus are distributed amongst several regulators. To further complicate this combinatorial regulation, factors may act as either activators or regulators depending on the context of their regulation[160]. For example, HB itself is known to both activate and repress depending on cofactor availability and concentration[161]. The tree models for each locus need not assume that a regulator has a strict role as activator or repressor, but rather can infer context-specific roles for each factor in each partition. In the mid-blastoderm *hunchback* tree model (Figure 5), the tree predicts that *hunchback* is activated in the anterior by three factors: BCD, Dichaete (D), and FTZ. Similarly the model predicts several repressors of *hunchback* expression: Huckabein (HKB), KR, GT, and KNI. These specific predictions are then testable by elimination of the proposed regulator and measurement of *hunchback* expression. In the case of *hunchback*, repression by each of the predicted repressors is documented by previous work on the gap gene network and cross-repression between its members, as is activation by BCD[162]. However, the proposed role of FTZ as an activator of *hunchback* expression in both the trunk and posterior domains is a novel prediction.

Previous work dissecting the *hunchback* regulatory regions showed that posterior repression was chiefly the role of HKB. The same work proposed that while BCD was the principle activator of the anterior regulatory region, separate regulators were responsible for driving expression in the posterior[162]. The authors show that the posterior expression can be generated in a reporter model with sequence containing several binding sites for TLL, consistent with the early blastoderm tree predictions. However, the authors also note that *hunchback* expression is clearly more complicated than this, as evidenced by two observations: 1) in *tailless* mutant embryos, the posterior *hunchback* stripe is still weakly expressed, and 2) the anterior cap and stripe patterns are generated independently and the same sequence that generates the posterior expression in a reporter construct also drives expression of the anterior stripe, but not the anterior cap pattern. In the mid-blastoderm tree model, FTZ is predicted to activate both in the posterior region and the anterior stripe.

To test this prediction, I co-stained early and mid-blastoderm embryos of both *tll* and *ftz* genotypes with mRNA hybridization probes for the *hunchback* gene product and antibodies raised against either the FTZ or TLL proteins. As found by Margolis, et al, functioning TLL was required for proper expression of the posterior domain of *hunchback* (not shown). The effects of the *ftz* mutation are at most quantitative modifiers of the strength of the *hunchback* pattern. Nonetheless, it appears that *ftz* embryos, as assessed by the absence of FTZ in the early blastoderm, generate a similar, but delayed *hunchback* pattern in comparison to the wild-type embryos (Figure 6). That is, the resolution of the anterior pattern into a cap and stripe

appears to lag in *ftz* embryos when embryos at a similar progression of membrane extension are compared. Whether or not the *hunchback* pattern fully recovers to the quantitative levels is unclear, but qualitatively, it appears that the *hunchback* pattern does fully recover in the anterior and posterior (not shown). Thus, there is clear evidence that identifying *tll* as a direct regulator of *hunchback* expression in the posterior, supporting the work of Margolis. The evidence for *ftz* is less strong, given the lack of quantitative microscopy to confirm or dismiss the perturbation of the eventual *hunchback* pattern. If the principle effect of *ftz* is the timing effect shown here, it is almost certainly an indirect effect of FTZ protein activating other factors that repress the continued activation of *hunchback* in these anterior regions. While it is possible that the effects of both *tll* and *ftz* are partly derived through the trans-network, both proteins bind *hunchback* regulatory sequence as measured by BDTNP ChIP-Chip data. Binding sites for both proteins are prevalent in the regulatory sequence, suggesting that this binding is not spurious [Fig 7]. The binding of TLL overlaps the bound regions of both BCD and FTZ, suggesting that it can activate transcription from both enhancers. One TLL site precisely overlaps a strong FTZ binding site, allowing the possibility that the same site codes for the recognition of both activators at different stages development.

Challenges to Describing Pair-rule Patterns

The description of pair-rule patterns by the tree models often fails to recapture recognized regulatory relationships. For example, the early *even-skipped* pattern is more similar to a gap expression pattern, and the tree model readily recognizes the two established activators of this pattern as HB and BCD (not shown). However, at mid- or late-blastoderm stages, the *even-skipped* pattern is predicted almost entirely by other pair-rule genes. While it is thought that the pair-rule genes do cross-regulate each other at this stage of development, the tree model is unable to parse the deep autocorrelation of the data in these stages. This in essence reveals two fundamental properties of the linear model at the heart of the classification trees that do not limit the ability to describe the earlier gap patterns. First, the model cannot account for auto-regulatory mechanisms, such as are known to exist in *ftz* auto-activation or *eve* auto-repression[163-165]. Second, the model infers structure in the data based on covariance alone, with only the ability to use separate time points as a proxy for causation. Thus, where there are multiple regulators all tightly correlated over a space in the embryo, the model cannot readily discern which is most closely related to the pattern it is trying to describe.

In an effort to limit the space for auto-correlation in these predictions, trees for the pair-rule patterns were generated with only the maternal and gap factors used as predictors. In this case, some aspects of *even-skipped* regulation are described, including the repressive role of GT in the anterior, and some tenable hypotheses are derived, such as a role for *Dichaete* as an activator in the trunk, which is supported by binding data. This tree also depicts the interplay of the DV system, with *zen* repressing the level of *even-skipped* expression in the DV axis. However, the tree

fails to capture many other aspects of established *even-skipped* regulation, such as repression by other gap genes such as *Kruppel* and *knirps*.

Discussion

This work aims to infer regulatory logic of the embryonic transcriptional network en masse, by simultaneous computational dissection of the embryo into similar regions and description of the co-expression values of over one-hundred patterned genes across several time points in those regions. The classification tree models are able to segment the early gap expression patterns of the embryo into contiguous and biologically meaningful partitions of the embryo. In many cases, these models are consistent with established regulatory relationships in the early embryo. These relationships are quantitative, allowing for inference of different roles depending on the concentrations of regulators such as the BCD morphogen and HB, which is known to both activate and repress in a concentration-dependent manner. Also, the model is implicitly conditional, which allows for the factors to take on different regulatory roles depending on the context of other factors expressed in different partitions of cells in the embryo. For example, it is known that KNI is able to both activate and repress depending on which cofactors are present[161].

The gap patterning models generate specific hypotheses regarding which factors act as repressors or activators in specific partitions of the embryo. These predictions may represent indirect interactions of the transcriptional network, though the compendium of ChIP binding data collected by the BDTNP is considered in the model, suggesting in the case of FTZ or TLL activation of *hunchback* or Dichaete activation of *even-skipped*, that these regulatory relationships are direct.

In the case of *hunchback* patterning, the model accurately describes the earliest activation of *hunchback* by the BCD morphogen, and subsequent activation and repression relationships as the pattern develops from a simple anterior domain into three distinct domains. These segments of the embryo are readily partitioned into spatially contiguous *hunchback* expression domains by the classification trees that describe them, and the regulatory relationships for the BCD and TLL activators, as well as all the known repressors of *hunchback* expression are accurately described in their respective roles. Additionally, the models predict a time-dynamic tradeoff of the gap gene network to the pair-rule network as TLL activation in the posterior cap of *hunchback* expression shifts to activation by FTZ. This prediction provides an explanation to what Margolis and colleagues referred to as an “additional activity” in their earlier dissection of the *hunchback* regulatory sequences, wherein expression by the posterior enhancer of *hunchback* included the anterior stripe domain of *hunchback*, but not the anterior cap. This expression domain is predicted to be coordinately activated by FTZ along with the posterior cap.

Description of the later pair-rule patterns was not as successful as the description of the gap gene patterns. The classification model depends on linear models of the relationships between regulator and target genes, and non-linearities are accommodated inherently by the conditional branching of the tree model. For

example, the non-linear relationship of *bicoid* to *hunchback* is accurately and readily described by the tree model for early *hunchback* activation. However, higher-order non-linearities, including cases of auto-regulatory feedback as in the case of several pair-rule genes, present a challenge for these conditional linear models. To the extent that factors merely covary in the transcriptional network without effecting any regulatory impact on their covarying gene partners, the model will mistake this mere covariation for a regulatory relationship. The model does pare away potentially spurious relationships by considering the evidence for direct regulation from the BDTNP ChIP datasets, but binding in these datasets does not reflect *bona fide* regulation due to both high false-positive rates and the prevalence of real but non-functional binding in the genome.

The tree models are apt to describe simpler patterns rather than the more complicated patterns that emerge later in development, but it seems the simplicity of the regulatory architecture, not the pattern itself, is the determinant of how accurate the tree models are. For example, the *hunchback* pattern is accurately described in three separate domains, with interdigitated patterns from many genes, including pair-rule genes, coming between the three *hunchback* domains. However, in the case of *hunchback*, there are two principal early regulatory elements, one for the anterior and one for the posterior. In contrast, the *even-skipped* locus contains at least five regulatory elements. In the case of the *hunchback* posterior enhancer, some expression is driven at the anterior stripe, suggesting that the description of the enhancers as anterior and posterior is a misnomer. The seven-stripe pattern of *even-skipped* has been dissected such that regulation of stripes 1 and 5, stripe 2, stripes 4 and 6, and stripes 3 and 7 are all separate[166-168]. Several further computational efforts, such as random forest sampling with prior weighting[169] and factor analysis for segmentation of similar regions prior to tree modeling, were made in an attempt to computationally discover regions such as the *even-skipped* stripes that are regulated by similar logic but expressed in non-contiguous patterns. Importance sampling of the random forest indeed confirms that the trees are making robust decisions in their classifications (data not shown), but did not yield improvement in the description of more complex patterns.

It may be that the definitions of discrete enhancers in the *even-skipped* locus are a false construct. For example, it is known that the stripe 3+7 enhancer also drives partial expression in stripe 2, despite being thousands of base pairs from the stripe 2 enhancer. Thus, it may be that a better model for the nature of enhancers is that gene expression is driven by a single joint model of all bases in the genome and their ability to contribute to the activation and repression of a locus. Surely proximity is a predictor of the most relevant bases, but further compartmentalization is a construct most useful to explain the activities of regions that were conveniently digested by restriction enzymes and then found to have regulatory activity. This notion echoes the thoughts of Arnosti's "billboard"[170,171] model for transcriptional regulation in the embryo and seems supported by the strengths and weaknesses of the tree models discussed here.

Though much of the information required to pattern the embryo is present in the matrix of covarying factors and their expression across the nuclei of the embryo, there are several sources of regulation that are not taken into account here. One such source is the direct ability of a covarying factor to regulate its target gene. This is accounted for with some number of our factors by the orthogonal measurement of ChIP-chip data from whole embryos, but this data is limited in several ways. First, the ChIP-chip data suffers from a relatively high false positive rate. Second, there is pervasive non-functional binding throughout the genome by each of these transcription factors. Third, these data are not resolved in either space (i.e. with cellular or regional resolution) or time (i.e. all stages measured in the microscopy dataset are combined into one sample bin for the ChIP data). However, even if these were not limitations of the ChIP data, we lack a defined model for predicting the impact of factor binding on the expression of a nearby gene. We do not know, for example, when a factor is likely to activate versus repress. We also have reason to believe that concentration of the factors determines their effects a in continuous fashion, but the thresholds relevant to the activities of each factor are unknown. Additionally, the expression of given regulators in a given nucleus does not ensure that the regulators will be able to bind at a given locus due to the local chromatin conformation at that locus. And if binding is permitted at the distal enhancer, it is unclear what conditions allow for the activity of the distal enhancer proteins at the basal promoter. Finally, even if RNA polymerase is successfully recruited to the basal promoter, it is unclear that the polymerase will successfully elongate and transcribe the locus given a certain set of distal regulators present at the enhancers of the gene.

Methods

Classification and Regression Trees

Each classification tree was generated with the RPART package in R[172] from preprocessed vectors of data using the Rpy2 module and additional Python code. The model for the classification tree was a linear regression predicting one target gene from the set of vectors of relevant candidate regulators. The residual was chosen as the complexity statistic to decide whether or not to attempt further branching at the leaves of each branch. Initially, the complexity parameter was varied between 0.01 and 0.10. A complexity value of 0.03 gives good results for the gap gene patterns, with higher values dismissing clearly meaningful partitions, and lower values leading to multiple branches from the same predictor. Importance sampling was conducted using the randomForest package from R[169].

Factor data were included in the set of predictor vectors if the given factor was a plausible regulator of the target gene being predicted at the time point under consideration. For example, the gap gene *Kr* was predicted by gap and maternal genes such as *giant, bicoid, caudal, hunchback,* and *knirps* (along with approximately 30 others) at the earliest time point. Whereas the classification tree for even-skipped pattern in the late-blastoderm considered all expression data that could not be ruled as plausible by the absence of binding in the ChIP-Chip datasets (see

below). In the case where both protein and mRNA data existed for a given factor, and both were of ostensibly high quality, the protein data were used and the mRNA data removed. In some cases, for example *zen* and *zen2* mRNA, the *zen* data was much better than *zen2*, and the *zen2* data were removed from the set of predictor vectors.

If BDTNP ChIP-Chip data existed for a candidate regulator, the genomic binding profile of the factor was considered in the predictions. Due to the high false-positive rate and the existence of pervasive non-functional binding in the ChIP-chip data, the data were used only to remove candidates if there was no evidence for binding in the 1% FDR binding dataset. Evidence was permissively defined as significant binding within 10kb of the TSS or TTS of the locus.

In situ hybridization

Embryos of the following stocks were collected for one hour and aged for 2 hours before fixation in 5% formaldehyde in heptanes for 20 minutes:

OreR

P{*ftz*/*lacC*}1, *ftz14*/TM3, *Sb1* (BDSC #5333)

Df(3R)*tll-e*, *ca1*/TM6B, *Tb1 ca1* (BDSC #5415)
cu1 tlll49/TM3, P{*ftz*/*lacC*}SC1, *Sb1 Ser1* ryRK (BDSC #7093)

Embryos were stored in methanol until pre-hybridization. A custom protocol was developed for the four-color stains required for simultaneous detection of *hunchback* mRNA and the protein of interest with the BDTNP image acquisition pipeline (which was previously limited to three color stains). Embryos of each relevant genotype were pooled during pre-hybridization as to minimize experimental variation across the samples. Each pool was then stained for *hunchback* mRNA in Coumarin, ftz::*lacZ* mRNA as a reference pattern in Cy3, either FTZ or TLL protein via antibody staining with Alexa-633 conjugated anti-Rabbit secondary, and the nuclear dye Sytox Green as according to the Protocol S1.

Protocol S1: Four-color in situ hybridization for BDTNP Pipeline

Rehydrate embryos from methanol – 2hr

The embryos are stored in methanol, but need to be rehydrated before we can perform any histochemistry. We use a "stepped" rehydration procedure rather than just rehydrating directly into PBT+TX. This is supposed to better preserve morphology, but is not absolutely necessary. After rehydration, aliquot 20-50ul of embryos into each 1.5ml tube for staining (15 ul of fly embryos in methanol will be about 20 ul when rehydrated, and this 20 ul volume is what you want per tube). Rehydrate only what you plan to stain, leaving rest in methanol for future use. This assumes that the fixed embryos have been stored in 100% MeOH.

1. Rock the embryos 5 min in 1:1 EtOH / MeOH
2. Rinse 2X with EtOH
3. Rock the embryos 5 min in EtOH
4. Rock the embryos 5 min in 50% EtOH / PBT+ TX
5. Rinse the embryos 2X with PBT+TX
6. Rock 4X 15 min with PBT+TX

Prehybridization – 3hr

1. Rock 10 min in 1:1 PBT+TX / hybe at room temperature
2. Incubate 10 min in pre-warmed hybe at 55C
3. Change hybe and incubate 45 min at 55C
4. Change hybe and incubate 1 hr 15 min at 55C

Preabsorb anti-Digoxigenin-HRP (if used)

1. Take an aliquot of embryos (20-50ul) into an eppendorf-tube
2. wash 3 times with PBT+Tx in room temperature
3. put embryos into 1ml 1% BSA in PBT+Tx
4. add 20µl of anti-Digoxigenin-HRP into the tube to generate 1:50 stock solution.
5. nutate over night in 4˚C

Hybridization – 30 min + overnight

1. Take aliquots of prehybridized embryos into eppendorf-tubes. If the volume of the embryos is 20 – 50µl, 200ul total volume with Hybe is enough, if the volume is 50 – 120ul, 300-500ul total volume is OK.
2. Warm the embryos in a heat block or water bath into 55C – 59˚C.
3. Aspirate hybe from embryos so that they are just barely covered in the tube
4. Dilute probe in 100ul hybe. (1:100 for DNP, 1:50 for DIG)
5. Denature probe secondary structure by placing dilute probe at >80C for 2-3 min.

6. Snap cool probe on fresh ice.
7. Quickly add probe solution to the prehybe'd embryos.
8. Incubate at 55C overnight (> 10 hours, < 48 hours)

Hot Wash – 3.5 hr

1. Rinse with pre-warmed hybe (55C)
2. Change hybe and incubate at 55C for 5 min
3. Change hybe and incubate 2X at 55C for 15 min
4. Change hybe and incubate 2X at 55C for 30 min
5. Rinse 3X with PBT+TX at room temperature
6. Rock 4X for 20 min in PBT+TX with Roche Blocking Solution (1:5 Roche in PBT+TX)

Incubate with first probe antibody – 2.5 hr

1. Incubate with anti-DIG-HRP or anti-DNP-HRP for 2 hr at room temperature
    a. anti-DIG-HRP should be preabsorbed, and used at a final concentration of 1:200 to 1:500
    b. anti-DNP-HRP should be used at a final concentration of 1:100

Wash first probe antibody 2 hr 15 min

1. Rinse 3X with PBT+TX
2. Wash 6X for 20 min with PBT+TX
3. Wash overnight with PBT+TX
    a. Overnight wash is essential for anti-DIG-HRP, but optional for anti-DNP-HRP

First Color Reaction – 1 hr 15 min

1. Aspirate PBT+TX, leaving 100µl embryos + buffer. If there are more than 100µl embryos, double all volumes.
2. For every 100µl embryos + PBT-Tx, add 100µl Tyramide amplification diluent.
3. For every 100µl volume in tube, add 1µl Coumarin-tyramide and mix well.
4. Take an aliquot from each tube and place on microscope slide and cover with 22 x 22mm coverslip
5. Nutate tubes at room temp.
6. Observe the color reaction under the UV-filter on a fluorescence microscope. When a pattern becomes visible as brightness grains, stop the reaction by adding 1mL PBT-Tx. If there is not a pattern after 1 hour, the staining has probably failed. Stop the reaction at 1hour 15 minutes and continue to determine if it worked too weakly for the eye to detect.

Wash color reaction - 20 min

1. Rinse 5X with PBT-TX
    a. Embryos can be left at 4C overnight

Strip antibodies off embryos – 2 hr 30 min

1. Wash with 1:1 Hybe in PBT+Tx for 5min at 55C
2. Wash 4X 10 min with HybeB or Hybe at 55ºC.
3. Rinse 3X in PBT+TX
4. Wash for 15min in PBT+TX
5. Rock in 5% formaldehyde in PBT+TX for 20 min
6. Rinse 3X in PBT+Tx.
7. Rock 3X 5 min in PBT+TX
    a. Embryos can be left at 4ºC overnight
8. Rock for 30min in PBT+TX + Roche blocking reagent

Incubate with second probe antibody and primary antibody – 2 hr 15 min

If some of the first color reactions failed, take new aliquots to the microscope and inspect them. If there is no pattern even after the excess coumarin-tyramide has been washed away, discard the tubes that failed.

1. Incubate with anti-DIG-HRP or anti-DNP-HRP for 2 hr at room temperature
    a. anti-DIG-HRP should be preabsorbed, and used at a final concentration of 1:200 to 1:500
    b. anti-DNP-HRP should be used at a final concentration of 1:100
    c. primary antibody concentrations should be optimized for good signal

Wash second probe antibody – 2 hr 15 min

1. Rinse 3X with PBT+TX
2. Rock 6X 20 min with PBT+TX
    a. The first 3 of these washes shouldn't be longer than 20 minutes or the background will increase. If there is no time for further steps, the embryos can be left in the last wash overnight at 4ºC.

Second Color Reaction 1 hr 15 min

1. Aspirate PBT+TX, leaving 100µl embryos + buffer. If there are more than 100µl embryos, double all volumes.
2. For every 100µl embryos + PBT-Tx, add 100µl Tyramide amplification diluent.
3. For every 100µl volume in tube, add 1µl Cy3-tyramide and mix well.
4. Take an aliquot from each tube and place on microscope slide and cover with 22 x 22mm coverslip
5. Nutate tubes at room temp.

6. Observe the color reaction under the UV-filter on a fluorescence microscope. When a pattern becomes visible as brightness grains, stop the reaction by adding 1mL PBT-Tx. If there is not a pattern after 1 hour, the staining has probably failed. Stop the reaction at 1hour 15 minutes and continue to determine if it worked too weakly for the eye to detect.

Wash color reaction – 20 min

1. Rinse 5X with PBT-TX
    a. Embryos can be left at 4C overnight

Incubate with secondary antibody – 2 hr 15 min

This step is very similar to the primary antibody, but no need for additional blocking.

1. Incubate embryos for 2 hours on ice with an Alexa-conjugated secondary antibody (1:500 dilution)
2. Wash 6X for 20 min with PBT+Tx.

Staining nuclei – 30 min + overnight

1. Rinse 3X with PBT+TX
2. Bring volume to 500µl PBT+Tx.
3. Add 10µl Sytox green (1:100)
    a. Use separate tips for each tube. Because Sytox green is an intercalating dye, it should be treated as a potential carcinogen.
4. Mix sytox well by pipetting vigorously (don't shake), otherwise embryos will not stain properly.
5. Wrap tubes in foil and rock at 4ºC overnight or up to 48 hours.

Dehydration – 1 hr 15 min

1. Rock with 30% EtOH in H2O for 10 min
2. Rock with 50% EtOH in H2O for 10 min
3. Rock with 75% EtOH in H2O for 10 min
4. Rock with 87.5% EtOH in H2O for 10 min
5. Rinse 3X quick 100% EtOH

Mounting – 1 hr + 2 days for drying

1. Aspirate EtOH, add 75µl xylene per slide to be mounted to each tube
2. Wipe slides clean with EtOH, layout on paper towels
3. Make bridges using #1 cover slips
4. Pipette embryos up and down to keep them moving and add them to the slide.

5. Cover embryos with 350µl DePeX using cut P1000 tips. Use a different tip for each slide.
6. Pick a clean cover slip from EtOH with forceps, dry it with lens paper and drop onto sample. Beware of bubbles.
7. Allow slides to dry 2-4 days in dark.
    a. Even after drying, the slides should be kept flat for approximately one month, since the DePeX will flow slowly if the slides are sideways. The slides will be dry enough for staging embryos the following week.

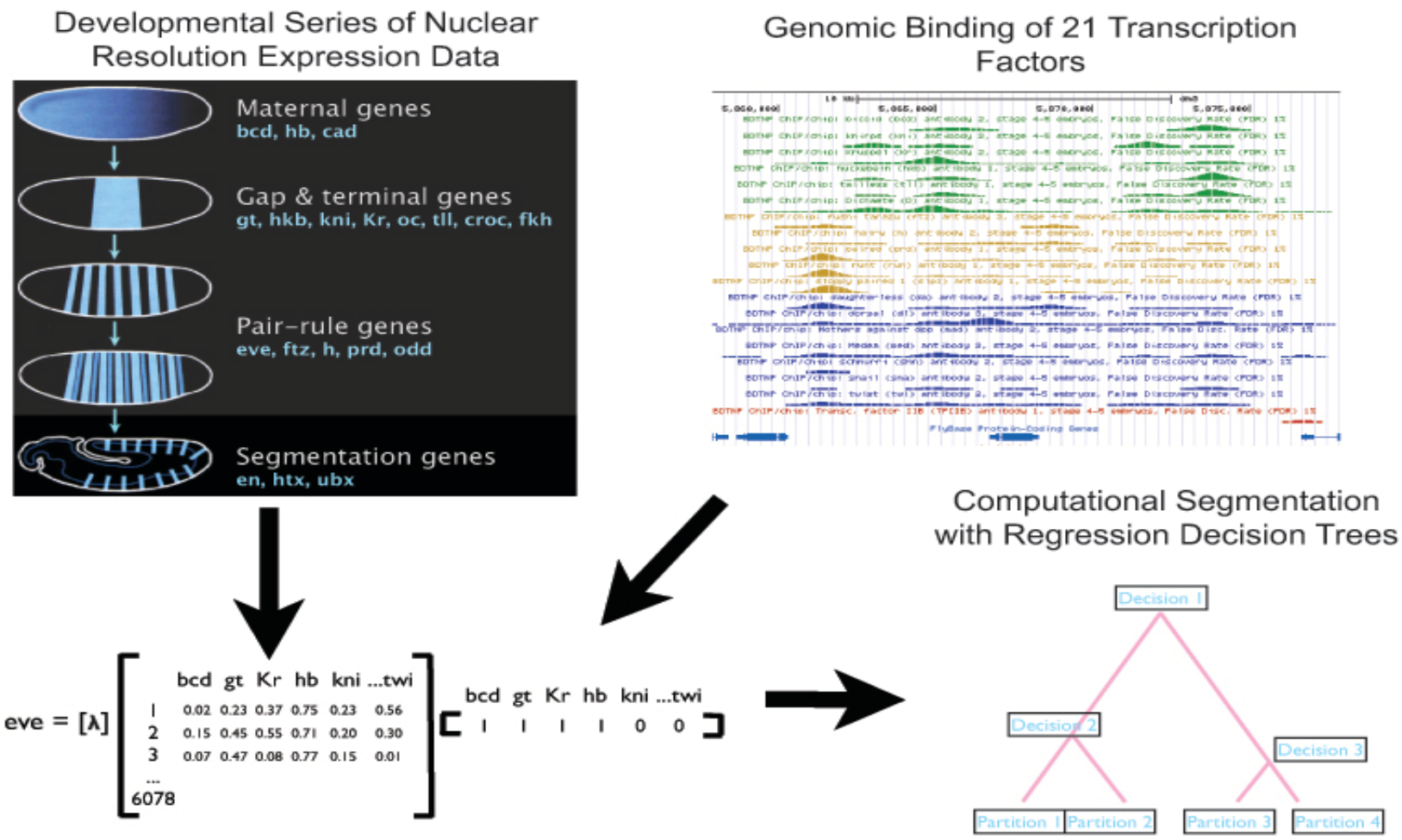

Figure 1: BDNTP gene expression atlas and ChIP-chip data are integrated into a single model for each target gene pattern for which a tree model is generated. The expression atlas is represented as a matrix of over 100 patterned factors in 6078 blastoderm nuclei, and the ChIP-chip data as a binary matrix where a regulator is effectively removed from the prediction matrix if there is no binding observed at the target locus.

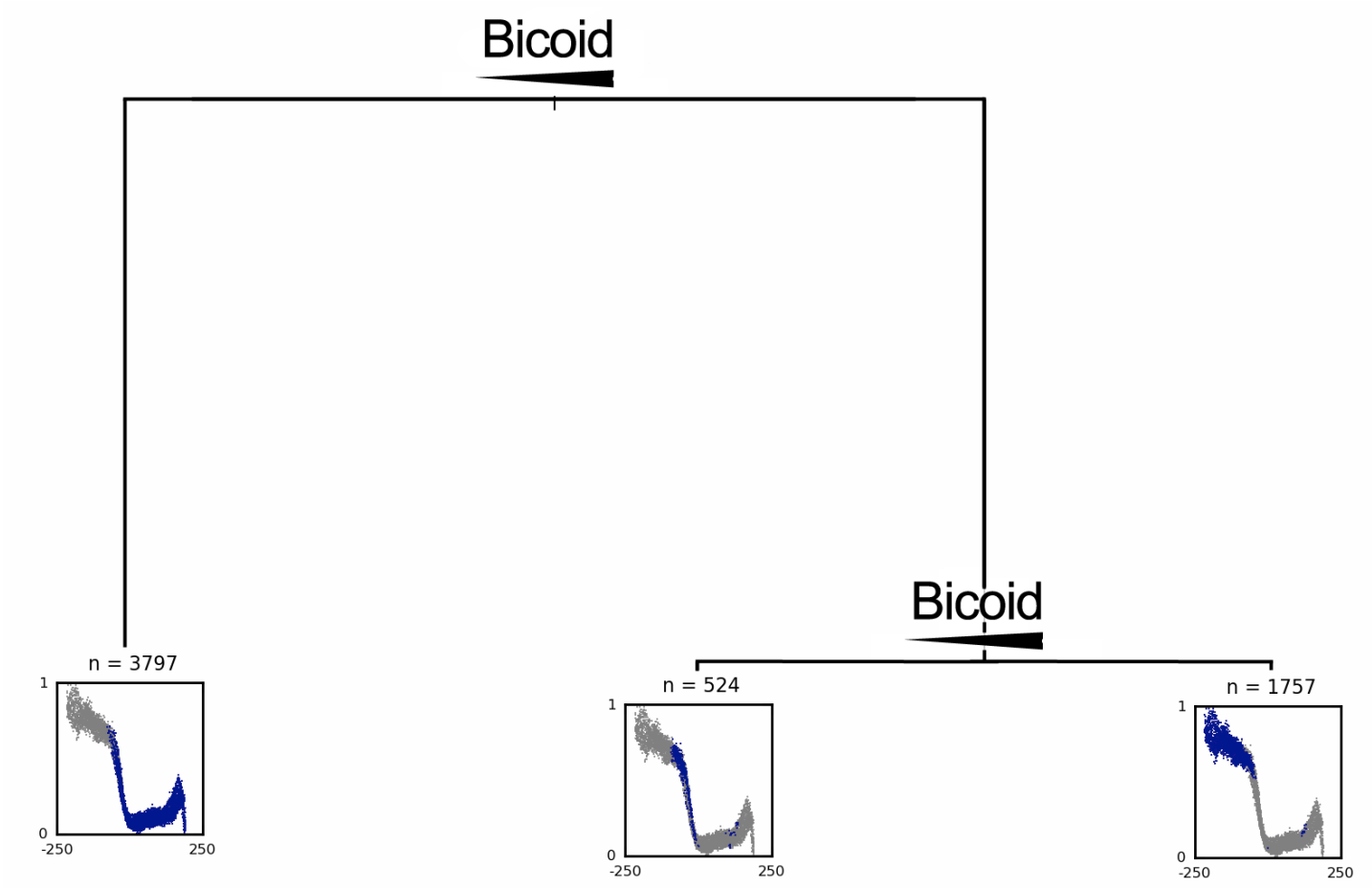

Figure 2: A classification tree describes three partitions of *hunchback* RNA expression at the onset of blastoderm cellularization (time point 1). At each branch point, the partition predicted by the lower level of the determining factor is branched to the left. At the leaf positions, the quantitative value of *hunchback* RNA is plotted for each nucleus from the anterior to posterior of the embryo (x-axis), with the arbitrary scaling of RNA concentration represented on the y-axis. Each leaf plot contains 6078 nuclei, with the nuclei belonging to that partition colored in blue.

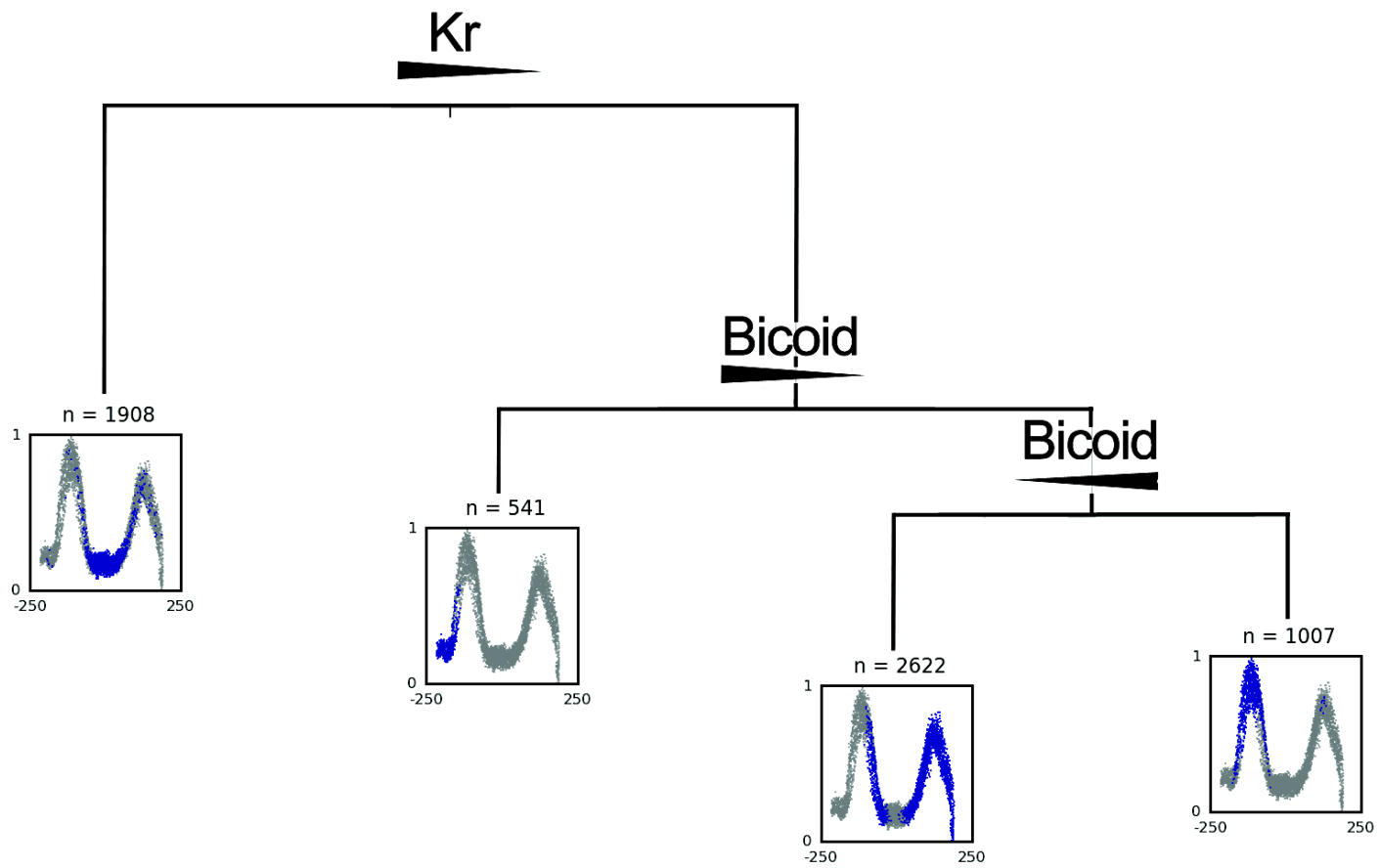

Figure 3: A classification tree describes the *giant* expression pattern at the onset of cellularization. This tree accurately captures known repression by the KR repressor protein, but also exhibits two errors owing to imperfections in the gene expression atlas. The anterior-most partition of 541 nuclei are predicted by a very high BCD protein concentration, but the atlas does not contain data for the TOR protein that inactivates BCD protein, and thus the tree model suggests BCD to have a repressive role. In the posterior-most cells, the atlas fails to suggest the CAD protein as the posterior activator of giant, and this is in part to subpar data for CAD protein at this time point.

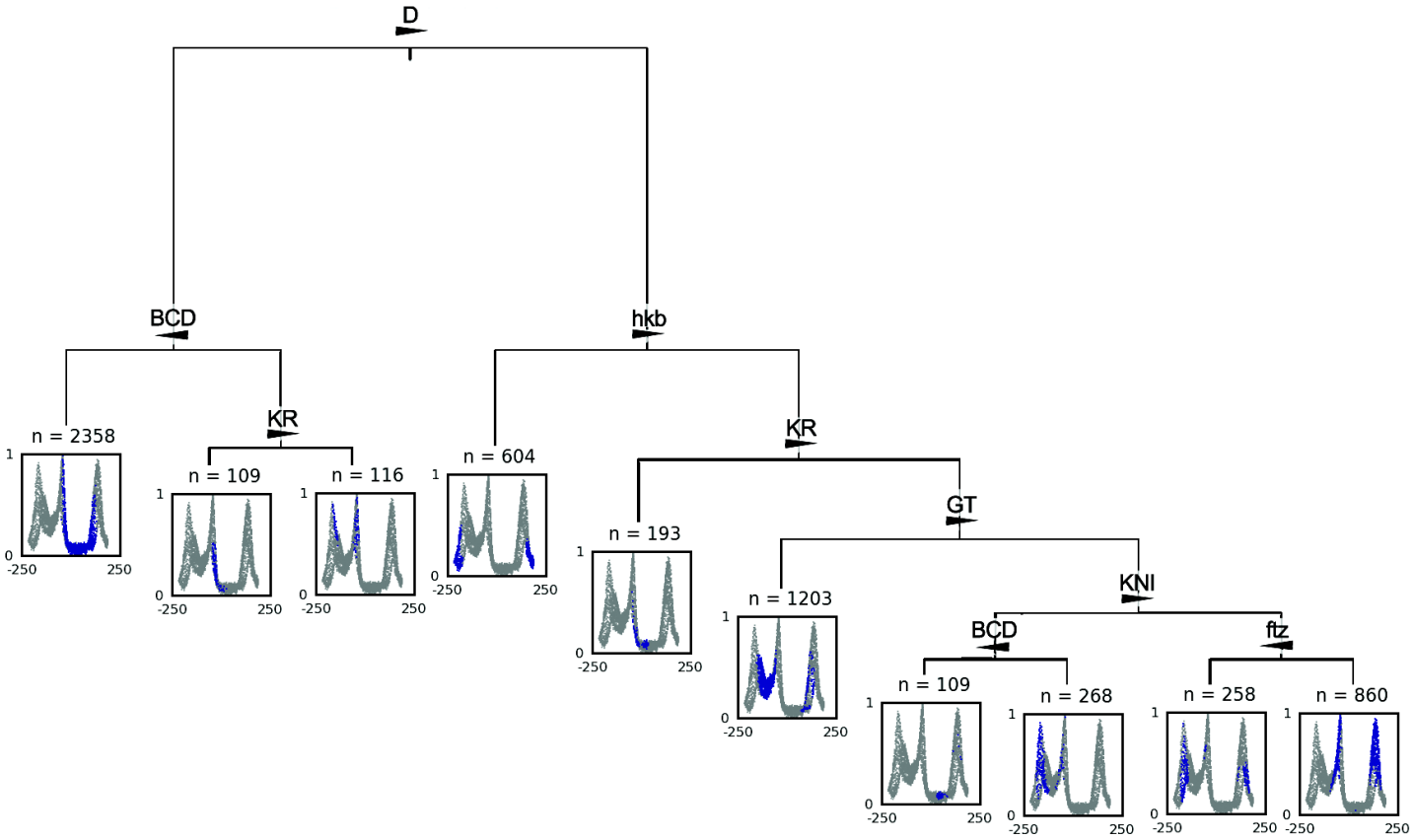

Figure 4: A classification tree describing *hunchback* expression in the mid-blastoderm after the activation of the pair-rule genes reflects the temporal dynamics and increased complexity of the transcriptional network.

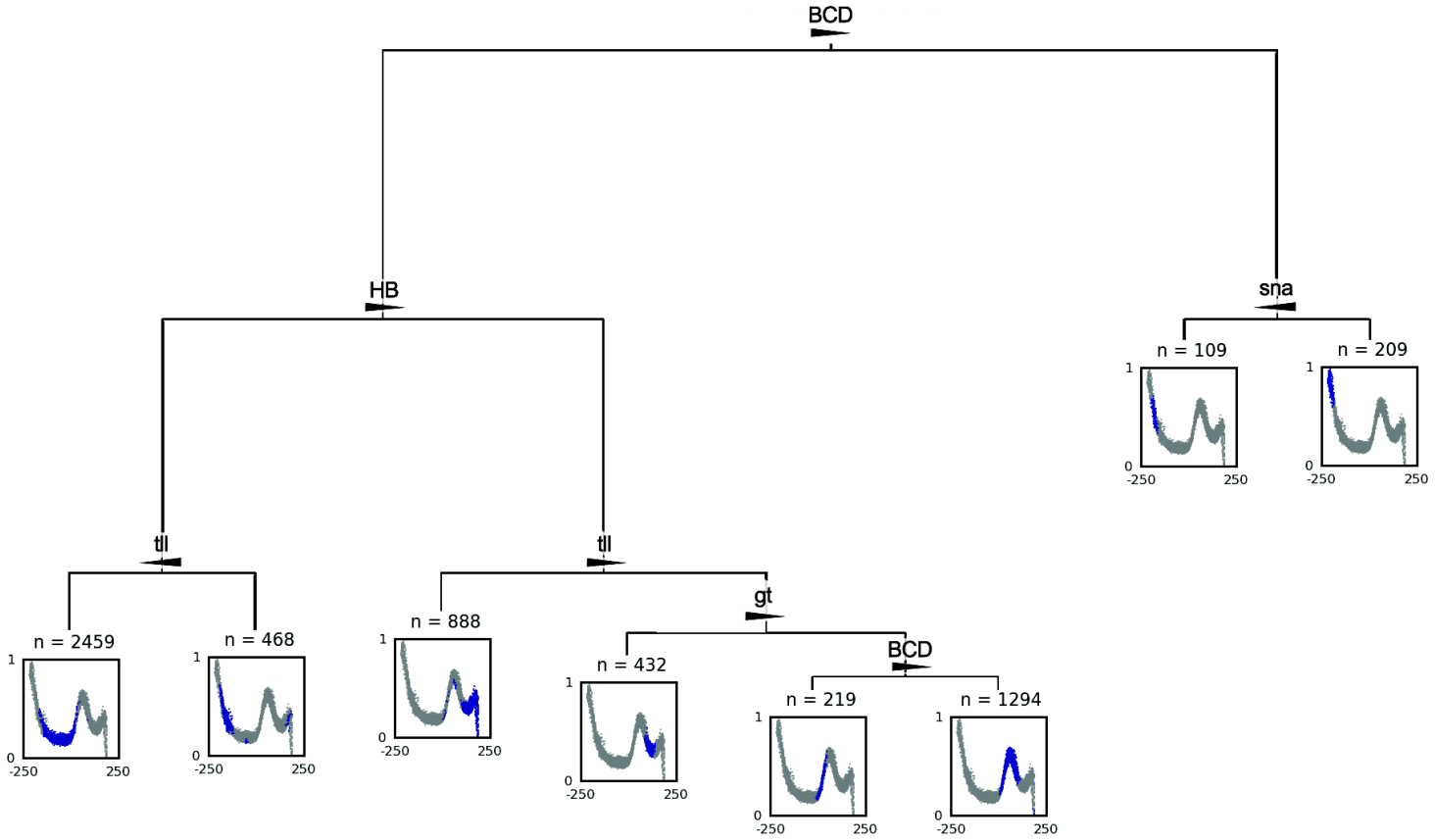

Figure 5: A classification tree for the early *knirps* pattern demonstrates the ability of the tree models to describe information from both the anterior-posterior and dorsal-ventral axes. RNA of *snail* partitions anterior-most nuclei into dorsal and ventral leaves.

Figure 6: ***ftz*** **mutants (red) appear to effect the posterior** ***hunchback*** **at most in a quantitative fashion, but separation of the two anterior domains is delayed in the** ***ftz*** **mutants (note two the anterior spikes in the WT trace (blue).**

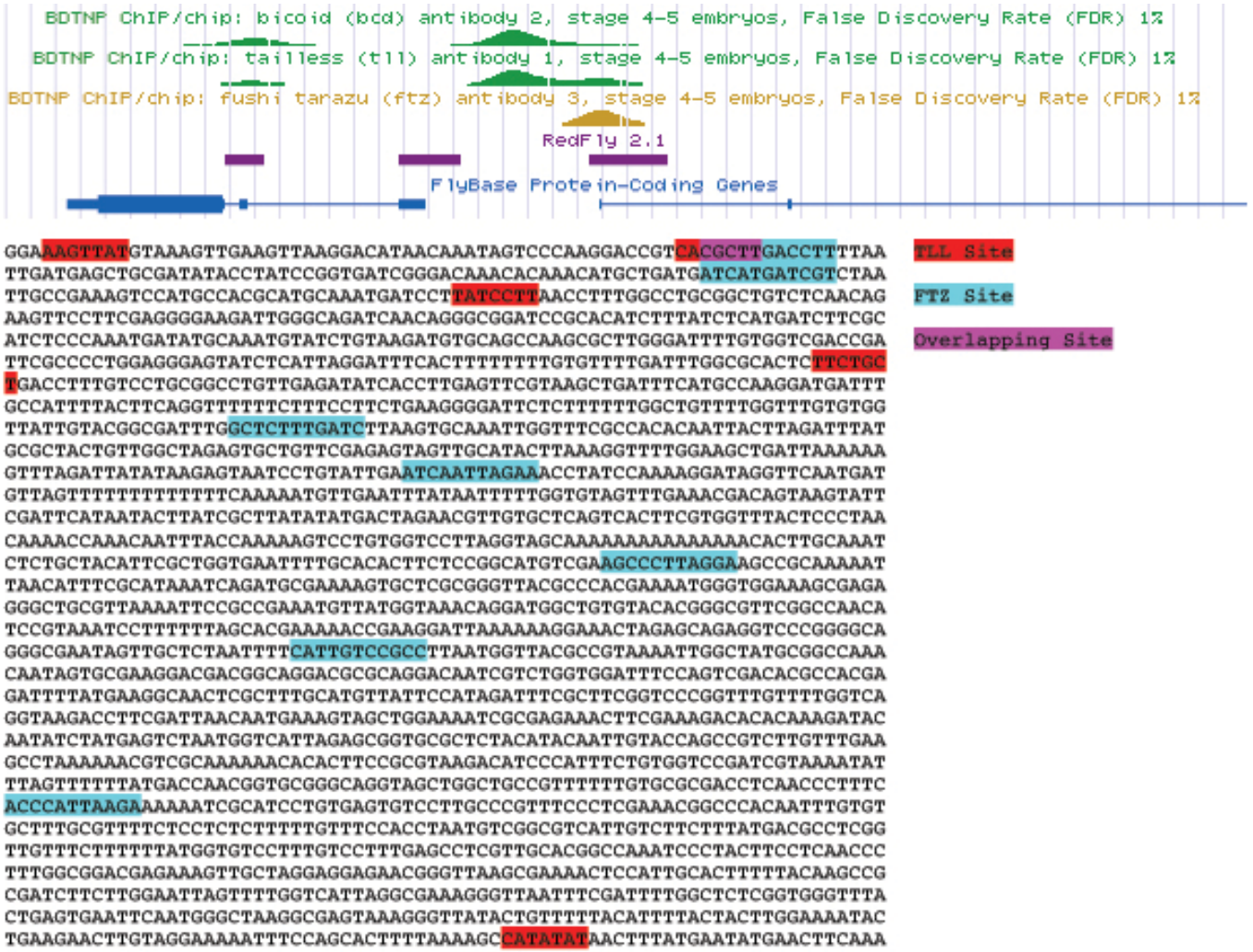

Figure 7: In addition to BCD, FTZ and TLL bind to the *hunchback* enhancer in the BDTNP ChIP-chip 1% FDR data. TLL binds overlapping regions with BCD and FTZ. Both FTZ and TLL sites (p< .001) are plentiful in the bound region near the two TLL peaks, overlapping at one site.

Chapter 3

The *In Vivo* Titration of Transcription Factor Dosage Alters Genomic Binding Profiles and Target Gene Expression in the Drosophila Embryo

Introduction

The nature of gene switches

The great diversity of animal forms arises despite the overwhelming conservation of animal gene sequences. This diversity owes largely in part to the great diversification of the regulatory regions that govern the expression of nearby genes by encoding short sequences recognized by the binding domains of sequence-specific transcription factors[12,13]. These factors bind in a combinatorial and coordinated manner to these promoter-distal cis-regulatory modules (CRMs) and either facilitate or inhibit the recruitment of the basal transcription machinery to the promoter of nearby genes, and in this way, the CRMs act as "switches" to direct the expression of those genes. The characterization of DNA-binding interactions typically involves *in vitro* titration assays to select DNA ligand with high affinity for a given transcription factor or DNA binding domain[115-118]. Alternatively, *in vivo* assays can be performed with the assistance of a cross-linking agent such as UV irradiation or formaldehyde to fix the DNA-protein interactions in histochemically purified nuclear extract, which can then be assayed to characterize the DNA footprint of bound factor[31,32,121]. With the advent of microarrays and high-throughput sequencing, these assays are now subject the massive parallelization allowing both higher quality estimates for the affinities of DNA ligands for their binding domains, and for genomic localization of DNA binding by a given transcription factor[147]. In recent years, the systematic characterization of DNA binding has been pursued in several models, yielding rich data lending significant insight into the nature of genome regulation. Comparative studies of this nature have established that the regulatory information encoded in the CRMs is more conserved than the primary sequence of homologous CRMs betrays[173], but also that evolution of this regulatory information is a major force of animal diversification[24,174-177].

From this data also arose a challenge to predict binding and subsequent gene expression from the primary sequence of the genome alone. This challenge remains unmet for several reasons. A fundamental understanding of binding energetics and kinetics *in vitro* is progressing thanks to atomic resolution experiments and super-resolution microscopy[102], but our understanding of *in vivo* occupancy is still based on averages of large numbers of molecular interactions. The temporal dynamics of the combinatorial and coordinate binding of the transcription factors themselves has only scantly been described even at the resolution of progressing developmental stages. Thus, the cooperative binding of cofactors or antagonistic competition between factors is difficult to integrate. Further, the sequence affinities determined by both *in vivo* and *in vitro* studies typically assume that there is only one class of ligand with affinity for the assayed binding domain, and the exceptions to these studies have observed cofactor-dependent sequence affinities. Also still in confusion is the relationship of transcription factor binding to CRMs and the chromatin state at and around the CRMs. Only sequence-specific DNA binding domains have the theoretical capacity to interpret the information encoded in the sequence of the regulatory regions[86]. The sensitivity of the region to DNaseI digestion predicts the

availability of the region to be bound readily be transcription factors[127,178], but the dynamics of chromatin remodeling at CRMs remain poorly understood.

In *vitro* experiments demonstrate that there is a continuum of binding energies between a binding domain and its ligands, with imperfect sequence motifs still able to bind tightly to the protein domain. It is clear that additional binding sites can be added to recruit more factor to a regulatory region, as shown with the GAL4 binding site in yeast[81]. However, *in vivo*, imperfect sites are present in CRMs, but it is unclear that they contribute less to the binding of their factors than stronger sites, or if their contribution is according to a function similar to the Hill kinetics observed *in vitro*.

Similarly, it is unclear that the concentration of the transcription factor itself regulates gene expression in a continuous fashion. The sigmoidal "switch-like" functions describing both binding and gene induction by the GAL4 activator *in vivo* suggestion that indeed the regulatory "switches" in the genome operate like a binary switch as the concentration of transcription factors regulating the switch region exceed some threshold[82]. In microbes, the titration of increasing allosteric activator does indeed drive graded increases in expression across a population of microbes, but recent work suggests this population increase does not reflect the kinetics within individual microbes at their respective promoters[179,180]. The existence of morphogen proteins, which direct differential gene expression programs according to their local concentration, also suggests that regulatory regions respond variably to a continuum of their regulatory factors[181-183]. The argument for continuous sensitivity, however, argues that in some biologically relevant nuclei, the concentration of the protein must not be present in extreme excess of the number of available binding sites. It is unclear that this is or is not the case given both the scarcity of quantified *in vivo* concentrations of transcription factors and the difficulty in defining discrete binding sites in a non-probabilistic manner. In the cases that are quantified, the concentration of protein does radically exceed the number of bound regulatory regions, but the total number of high-quality sequence matches in the genome is not dwarfed by the number of protein molecules per nucleus[86,122]. Additionally, it may be that even the most canonical of morphogen proteins affect regulation indirectly via induction of feed-forward and feed-back cooperativity loops, and that their concentration past a binary threshold is irrelevant[184].

Ergo a few models exist regarding concentration of transcription factor *in vivo* and the consequences of varying that concentration. It may be that, as the "flip switch" model of the GAL4 model suggests, when a transcription factor is expressed, it is expressed to radical excess of its genuine targets, saturating the binding curve of the available sites (Figure 1). Alternatively, it may be that binding sites throughout the genome are not saturated by transcription factor, even under strong induction. An intermediate model would allow for saturation at some sites, while the same concentration of factor fails to saturate weaker sites. In this model, if *in vivo* concentrations are saturating, then high concentrations of transcription factor would not increase binding at high affinity sites, but would at other lower affinity

sites. This model consistent by the mechanisms thought to operate in morphogen gradients[185]. A generalization of this model includes the possibility that unsaturated sites in the genome are non-functional, allowing for the possibility that some regions are sensitive to transcription factor concentration with respect to binding, but as a rule, there is no consequence to gene expression as a result of this sensitivity. This could be because evolution has not needed to tune the specificities of regulatory sequence and changes in binding do not effect changes in the expression of the target loci. It is also not clear that the regions most highly bound by any particular transcription factor are the most functional region, though these properties are generally assumed to be positively associated. Robustness could also arise from the non-linear properties of the transcriptional network, in the fly, for example, the cross-repression of gap genes[130,186] and the auto-regulation of pair-rule genes[187-189]. Finally, it could be the case that the appropriate model for understanding the role of concentration varies between transcription factors.

The maternally deposited activator *bicoid* is a homeodomain-containing transcription factor responsible for directing the anteriorization of the fly embryo[190]. Its canonical morphogen activity suggests that its targets are sensitive to varying concentrations of the factor throughout the anterior of the embryo, which are highest in the anterior tip and recede to a minimum in the posterior of the embryo[44,191]. The decoding of the morphogen position and concentration has been the focus of much molecular experimentation and computational modeling, but is still not well understood[144]. In surveys of genomic binding by ChIP-chip and ChIP-seq, BCD protein binds to fewer regions than other early regulators of embryonic patterning in the fly[122,192], though still 5-10 times more regions than have been characterized by small-scale validation[184].

The zygotically transcribed gap gene *Kruppel* is a zinc-finger containing protein with well-characterized repressor activity[99,193,194]. Unlike *bicoid*, it is expressed in one central domain, which is tightly contained by flanking gap gene repressors both the anterior and posterior sides[195,196]. Genomic binding data for *Kruppel* identified an order of magnitude more high-confidence bound regions than for *bicoid*, though the number of small-scale validations for KR protein regulation are considerably fewer than for BCD protein[122,147].

Here, I have collected genomic binding data for both BCD and KR proteins in an allelic series of dosage variants for these proteins. The maternal genetics of *bicoid* allow for both lower and higher dosage of the gene product with existing fly lines, and for *Kruppel*, a copy number variant with twice the natural dosage of functioning *Kruppel* alleles was created. In addition to binding data, expression of the transcriptome was profiled with single-embryo RNA-seq[197], and microscopy samples were generated for dozens of target genes in whole embryos. These data show that varying gene dosage and protein concentration does have a considerable impact on genomic binding of these two transcription factors *in vivo*. As concentration is increased, binding increases at many sites with high primary sequence affinity for the protein. The putative function of these dosage sensitive

regions differs for each transcription factor, but the dosage sensitive regions are associated with differential expression of nearby genes.

Methods

Characterization of mutant flies

The allelic series of *bicoid* dosage variants consists of 1, 2, and 6 copies of the locus, which is transcribed by the mother and deposited to each offspring, no matter their genotype. The following flies were used in the allelic series:

bcd M12-3 (bcd/TM3,Sb)

OreR (wildtype)

bcd +5+8 (double P-element insertion on *X*)

The mutant *bicoid* flies were a gift from Stephen Small at NYU and are descendents of the flies created by Wolfgang Driever.

The *Kruppel* over-expression flies are the result of a homozygosed insertion of the bac recominbeered CH321-25N18[198] region to the attP2 site located on chromosome arm 3L.

Mutant expression for each line was confirmed by RNA-sequencing and antibody staining for the mature protein product. As expected, the expression pattern of BCD protein varies according to gene dosage, but the domain of KR protein expression is contained to the same domain as wild-type. Additionally, local protein concentration at fixed points in the BCD gradient varies according to dosage, as was observed previously[44]; KR dosage increase similarly confers higher concentration per nucleus.

ChIP-sequencing

Antisera was purified using constructs made available by the BDTNP for BCD and KR epitopes. The purified antibody was used at approximately 1ng/ml in immunopurification.

Mutant embryos were collected for one hour and aged to the early blastoderm stage at 25C before dechorionation and fixation. Nuclei were purified and sonicated, and chromatin was immuno-purified according to Protocol S2. The resulting DNA samples were Illumina sequenced yielding approximately 5,000,000 mappable reads for each sample.

Reads were mapped with bwa64[199] to the dm3 genome made available by the UCSC Genome Browser. Read coverage for each sample was scaled to 10,000,000. Bound regions were called with MACS14[200], and for each factor, the wild-type *OreR* flies used as the reference set. The most significantly bound regions for each factor (FDR

< 1%) were compared across the samples, and where multiple samples overlapped, the regions were trimmed to the intersection of the called regions.

RNA-sequencing

Two replicates of single embryos for each mutant were collected and staged from nuclear cycle 10 through 14 and whole RNA was extracted. For stages relevant to *bicoid* and *Kruppel* expression, the RNA was poly-A selected and Illumina sequenced, generating on the order of 30,000,000 reads per sample. Reads were mapped using the Tophat/Cufflinks[201] suite and differences in expression compared to wild-type data were determined by using the stage-matched raw data from *CaS* flies. Cuffdiff was used to determine fold change and significance with all samples from each dosage series considered jointly.

Primary Sequence Affinity

Primary sequence affinity was calculated by scanning the region of interest for matches to the probability weight matrix for the given factor[125]. Only significant matches (p<.001) were considered. The final score is exp(sum(-log(score pvalue))).

Microscopy

Embryos were stained and imaged according to Protocol S1 in Chapter 2.

Additional analyses

All additional analyses were performed with custom Python code. This code is available on request, and will be made publically available in the future.

Results

Dosage Series Mutants Constitute a Titration of Their Gene Products

The mutant fly lines in used in this study were characterized to verify that the increase in gene dosage produced an increase in local gene product concentration. The RNA levels for both *bicoid* and *Kruppel* lines were compared to wild-type levels and verified to constitute a dosage series. FPKM values for *bicoid* demonstrate a surprisingly consistent RNA dosage. 2X flies contained just slightly less than twice the RNA of 1X flies, and 6X flies contained slightly less than three times the RNA of 2X flies (Figure 2). Protein levels for the *bicoid* series were established previously, but to confirm this both *bicoid* and *Kruppel* flies were antibody stained and await quantification by the BDTNP microscopy pipeline.

Many Bound Regions are Sensitive to Gene Dosage

To compare relative binding between the samples, the scaled read density was calculated across each bound region, and the regions ranked accordingly for each sample. Absolute binding is unknowable from these data, as a decrease in binding for a particular region in one sample does not necessarily mean that there is less

binding at in one sample versus the others, as it could also be the case that all other regions increased while the particular region stayed constant. Nonetheless, the regions that shift position in the rank list comparison can be interpreted as regions that were bound relatively differentially across the samples. Signal to noise ratio varies among the samples, which is attributable to variation in IP efficiency and library construction, but samples with the best signal were selected from each genotype for the analysis.

Variation in *bicoid* dosage between the samples clearly disrupted the rank order of 392 bound regions in the 1% FDR set (Figure 3A). Reduction of BCD had a larger effect than increasing BCD, though the top of the ranked list was largely invariant for all three samples. The disruption of binding across 1770 bound regions in the KR ranked lists was less pronounced than either BCD variation (Figure 3B).

Functional BCD Bound Regions are Insensitive to Dosage Variation

The widespread disruption in BCD binding demonstrates that some bound regions are sensitive to BCD dosage. However, the most highly bound regions were also highly consistent across the dosage series. All of the regions considered in the ranked lists are likely to be genuinely bound *in vivo* with the FDR cutoff set at 1%. However as non-functional binding is a pervasive trait of the transcription factors in the fly blastoderm[122,202], I asked if these regions were enriched for, or were proximal to, *bona fide* functional regions. The distance of bound regions to validated blastoderm CRMs from the RedFly database[203] is zero or close to zero for the regions at the top of the ranked lists in each sample (Figure 4A). Though there are varying signal to noise profiles in the three samples, this measure confirms that the majority of *bona fide* regions are recovered in each sample.

To ask if there exists a difference in the regions biased toward higher binding in higher or lower protein dosage conditions, difference-ranked lists were compared for the distance of each bound region to a RedFly region. RedFly regions, plotted as the array of data points at zero on the y-axis, are enriched at the high-consistency regions, near zero on the x-axis (Figure 4B and C). Additionally, many regions that are close to, but not directly overlapping, known RedFly regions are also exhibit high consistency between the samples. In both 1X and 6X dosage flies, the regions with relatively higher binding are farther from RedFly regions than are regions more highly bound in the 2X flies. This could be due to signal variation in the datasets, or to a real effect of increased protein binding to additional non-functional sites in the 6X case, and real reductions in binding at functional sites in the 1X case.

KR Bound Regions are Not Enriched for Functional Annotation

Most functionally annotated RedFly blastoderm CRMs are represented in the KR bound regions. However, unlike BCD bound regions, KR bound regions in both samples are not enriched for functional annotation generally, and only minimally at the very top of the ranked list (Figure 5A). Consistent with this observation is the lack of proximity to RedFly regions in the high-consistency regions (Figure 5B),

however the distribution of RedFly regions appears symmetrically distributed about the center of the ranked difference list, suggesting both samples recovered the RedFly regions with similar efficacy. There is a noticeable flattening of the LOESS[204] curve at the high-consistency regions, suggesting that the high-consistency regions are of a similar and intermediate distance to the nearest RedFly regions. If there is any trend in this data set regarding bias toward one set of bound regions, it is for closer proximity in the 4X sample, but it is less pronounced than for the BCD samples.

Primary Sequence Affinity Corresponds to Dosage Sensitivity

Higher BCD dosage increases relative binding at many regions, and these regions are enriched for high primary sequence affinity (Figure 6). For BCD samples, the regions biased toward relatively lower binding in the 1X dosage sample have symmetrically lower binding than their 2X biased counterparts. Thus, the BCD bound regions are sensitive to both higher and lower BCD concentrations. However, the regions biased toward higher binding in higher dosage samples are not the highest affinity regions. The high-consistency regions have the highest sequence affinity of all BCD bound regions, while the regions biased toward lower dosage samples exhibit lower primary sequence affinity. The sequence affinity of 4X biased regions in the KR data are of the highest sequence affinity, asymmetrically higher than their 2X biased counterpart regions. In contrast to BCD samples, the KR high-consistency samples are of similar and intermediate sequence affinity (note the flattening of the LOESS curve in Figure 6D).

Changes in Binding Correspond to Changes in Gene Expression

Dosage sensitive variation in binding is associated with variation in the expression of gene nearby the dosage sensitive bound regions (Figure 7). Dosage sensitive regions biased toward increased binding in the 6X BCD sample were expressed at higher levels than their 2X biased counterparts in genomic expression data (Figure 7A). If statistical significance criteria reported by the multivariate model of CuffDiff (see methods) are applied to these results, the data are more sparse, but the association strengthens. Similarly, regions biased toward increased binding in the KR 4X sample exhibit lower expression in the genomic expression data (Figure 7C). These results are consistent with the well-established roles of BCD as transcriptional activator and KR as transcriptional repressor. In the case of BCD under-expression, there is not a predominant trend in gene expression, as 1X biased regions do not generally exhibit lower expression as compared to the high-consistency or 2X biased regions (Figure 7B). This is consistent with the flatness of the LOESS curve describing the sequencing affinity in the 2X biased regions relative to 1X biased regions in Figure 6C. Interestingly, the high-consistency regions in all samples express at similar and intermediate levels, as illustrated by the flattening of the LOESS curve in each subplot near the center of the x-axis.

Changes in Relative Binding are Present at Some Functional BCD Bound Regions

While it is generally the case that functionally annotated BCD bound regions exhibit more consistency in their binding and gene expression profiles, there are clearly cases of divergence. Figure 8 shows three cases of *bona fide* BCD targets near bound regions with variable binding profiles. In the case of *btd*, an anteriorly expressed gene with a dosage insensitive expression pattern, there are two binding site clusters of high sequence affinity (Figure 8A). The binding to the CRM is very consistent for 2X and 6X samples. The 1X sample is bound more strongly at the 5' cluster relative to the 3' cluster, which is the reverse of the 2X and 6X binding profiles. As was the case for the genomic data set, the stronger binding site cluster is bound relatively more strongly in the presence of BCD concentration. Expression of *btd* is very similar across the samples, but while the expression domain constant across BCD dosage conditions, the FPKM of the sample does modestly increase in the 6X dosage.

The gap gene *hunchback* is the earliest target of BCD activation in the embryo, and it is critical for setting the anterior compartment boundary of the embryo. The expression domain of *hunchback* is very sensitive to BCD dosage, extending nearly 50% egg-length from 1X to 6X dosage. Accordingly, the FPKM of *hunchback* rises more than two-fold in across the dosage series. Binding of BCD to *hunchback* is present at two well-characterized CRMs[162,205] (Figure 8B). BCD protein binds strongly to both of these regions, and binding to both is required for proper expression of the *hunchback* pattern, though binding at either is sufficient for anterior expression. Despite the co-functionality of these CRMs, only one of these regions appears to be sensitive to BCD binding, even though both regions have very similar primary sequence affinity. The distal enhancer is bound relatively more strongly than the P2 enhancer, but increasing BCD dosage increases the binding only at the P2 enhancer.

The expression pattern of the *eve* gene is BCD dosage sensitive, but the whole-embryo expression levels of the pair-rule gene appear consistent, though the 6X sample levels are modestly higher than 1X and 2X (Figure 8C). Like *hunchback*, there are two well-characterized *eve* CRMs activated by BCD. The strongest binding is at the stripe 1 enhancer[166], for which the relative binding profiles for each dosage are similar. At the stripe 2 enhancer[167], however, there is relatively stronger and very similar binding in the 2X and 6X samples. Similar to the *btd* case, the stronger binding site cluster is relatively more occupied.

These results also come with two caveats that merit repeating. First, the binding profiles compared here are relative within and between samples. Though a particular peak may have a similar binding profile at all dosages, it is possible that one sample contains proportional increase or decrease at all loci. The disproportionate change in binding of adjacent bound regions near the same gene is thus the strongest evidence for a genuine change in binding, but the absolute scale of binding in each sample is unknown. Parsimony suggests that a higher BCD Dosage should generate relatively higher binding in these cases, but it is not strictly true. Second, varying *bicoid* dosage variably anteriorizes the embryo, thereby

transforming the expression domains of many anterior genes, and these results should be considered with that in mind, but there are several lines of evidence reviewed in the discussion suggest that this is not the principal or exclusive phenomenon driving these observations.

Discussion

Transcription Factor is Not Saturating at High-Confidence Bound Regions

In this work, I have shown that many genomic regions bound with high-confidence by the canonical sequence-specific transcription factors Bicoid and Kruppel are sensitive to the dosage of these proteins. The changes in relative binding at these dosage sensitive regions are structured with respect the functionality, sequence affinity, and associated gene expression from nearby coding regions, however the structure in these dimensions varies between the two factors. Regions bound by increasing levels of BCD protein have a high affinity for the BCD protein, but not as high as the regions with consistent binding between the samples. These high consistency regions are dosage insensitive with respect to binding and to the association with expression from nearby genes. The highly consistent KR bound regions are not the enriched for functionally annotated regions, nor are they the most highly bound regions. However, regions sensitive to the increased dosage of KR have the highest sequence affinity for the protein of all KR bound regions. In the case of both BCD and KR dosage sensitive regions, where the transcription factor is at relatively higher occupancy, the expression of nearby genes changes are consistent with the roles of these two proteins as activator and repressor, respectively.

These results suggest that these canonical transcription factors are not present in concentrations that radically exceed the number of available high affinity binding sites in the genome. However, the bulk of the changes do not appear in either case to be predominantly at functionally annotated regions. This supports a model of quantitative binding and quasi-functional regulation of many sites, where binding and nearby gene expression can be influenced by the concentration of a transcription factor, but the most functionally important sites are insensitive to quantitative variation in the factor concentration.

However, some changes are present at canonical and functionally annotated targets of the BCD transcription factor. Though the changes in binding and gene expression observed are no doubt in part due to the variable anteriorization of embryos in the dosage series of BCD protein, several lines of evidence argue that this is not the only force underlying these observations. First, I observe disproportional changes in binding at adjacent bound regions both known to functionally require BCD for proper activation of their target locus. If additional binding at the *hunchback* CRMs was the product of expansion of the *hunchback* expression domain with higher BCD dosage, then each additional BCD-expressing nucleus should proportionally increase binding to both *hunchback* CRMs, canceling in terms of their relative binding profiles. Second, some targets of BCD do not appear to have dosage sensitive

patterns, such as *btd*. Yet at this locus, there is disproportional increase in binding at the 3' binding site cluster of the *btd* CRM. Finally, the sequence signal associated with increased binding of BCD suggests that it is indeed the case that most protein is bound to the sequence with highest affinity, and there is no reason to expect this result with the trivial model of expanded anteriorization. That the same relationship of strong sequence affinity and higher relative binding is observed for KR suggests that this model phenomenon can be generalized, reinforcing the observation.

Bound Regions Sensitive to BCD and KR Dosage Behave Differently

Though both BCD and KR dosage sensitive regions share enrichment for high sequence affinity and association with nearby gene expression, they differ with respect to their overall binding, binding consistency across samples, and enrichment for functionally annotated regions. The KR bound regions with the highest overall binding in both 2X and 4X samples do not have the highest affinity for KR, though generally speaking of the 1% FDR bound region list, sequence affinity for KR is high. This is in contrast to BCD, which exhibits the highest sequence affinity at highly bound and functionally annotated regions, which are also not enriched in the most highly bound KR regions. It may be that for KR, the highest sequence affinity has not been selectively tuned for binding at the functional sites. Perhaps a better understanding of the biophysics of binding by this C2H2 zinc finger protein would reveal a reason why suboptimal binding is best for functional KR bound regions. Alternatively, it may be that KR functions as more than a repressor at CRMs that direct expression of embryonic patterning genes. The definition of functional region here is an annotation of blastoderm CRM function, but it is conceivable that KR is bound more pervasively to targets in the genome distal from blastoderm CRMs than BCD because it has a broader role in repression or chromatin insulation. Comparative data for KR binding may shed light on these alternative models, if pervasive binding at the highest affinity sites distal to blastoderm CRMs is conserved between species of *Drosophila*.

A Model for In Vivo Occupancy and Kinetics

The data collected suggest that both BCD and KR bound regions are sensitive to dosage. Notably, in the BCD samples, there is evidence that the regions are responsive to a lower dosage with respect to binding, though nearby gene expression seems generally unaffected. This suggests that while heterozygosity of a transcription factor is broadly sufficient for wild-type gene expression levels, it also suggests that there is excess BCD protein and excess BCD binding present in the wild-type embryo. This is consistent with the observation that it is the highest affinity sites that absorb increasing amounts of protein, raising their occupancy. This is in contrast to the model provided in Figure 1B, which suggests the strongest affinity sites will become saturated first, possibly with weaker sites absorbing excess protein thereafter. While the model from Figure 1B seems true *in vitro*, and has some support from previous study of the *Dorsal* morphogen gradient, the data here suggest that there are copious high affinity bound regions with capacity to bind increasing molecules of BCD protein.

However, these dosage sensitive sites are generally not *bona fide* functional bound regions. That the BCD functional sites are generally dosage insensitive, but of comparable or even higher sequence affinity compared to the dosage sensitive regions suggests that these sites may indeed be saturated. It also suggests forces outside of factor concentration and primary sequence affinity are determining the affinity of these sequences. For example, the *eve* stripe 2 enhancer has considerably higher primary sequence affinity for BCD, but binding in all samples is much lower than the binding at the stripe 1 enhancer. The binding domains of BCD and KR are known to directly compete for overlapping sites, in addition to mechanisms of short-range and long-range repression attributed to KR. KR binding is considerably higher at the stripe 2 enhancer. Perhaps the stripe 2 enhancer exhibits dosage sensitivity not because of its higher primary sequence affinity for BCD, but because the two proteins are antagonistically competing at the enhancer, and the increased BCD concentration mitigates the repressive effect of nearby or directly-competitive KR binding. It is also possible that differing profiles of DNase I sensitivity[127,206] or binding of the BCD co-activator Zelda[207] may inform the mechanism underlying differential dosage sensitivity of these two enhancers.

Transcription Factor Concentration, Disease, and Robustness in Transcription Networks

It is worth mentioning that each of the mutants comprising the two dosage series here are viable flies. Though the *bicoid* mutant fly lines are both quite sick and proved difficult to collect samples from, this is may be due in part to the accumulation of deleterious alleles on the balancer chromosome of the 1X flies, and to the *CyO* allele carried by the 6X line. These flies and the *Kr* 4X line both produce viable offspring in standard development times. For the *bicoid* over-expression flies, it is known that massive cell death degrades the expanded anterior compartments, and that larvae develop normally. In both series, there is widespread disruption to the expression levels of the gap and terminal genes, but the pair-rule genes are expressed at remarkable consistent levels. Additionally, heterozygous *Kr* deficiency flies exhibit molecular patterning phenotypes (not shown). These observations suggest that the dosage effects of BCD and KR reverberate though the transcriptional network, but that compensation takes place prior to gastrulation, such that segmentation and tissue derivation proceed properly. These patterns are consistent with what is known of disease associated with transcription factor copy number variation. Despite the roles in directing the expression of dozens to hundreds of critical developmental genes, developmental transcription factors are rarely haploinsufficient, perhaps owing to the robustness conferred by the buffered kinetics and occupancy at functional sites observed here.

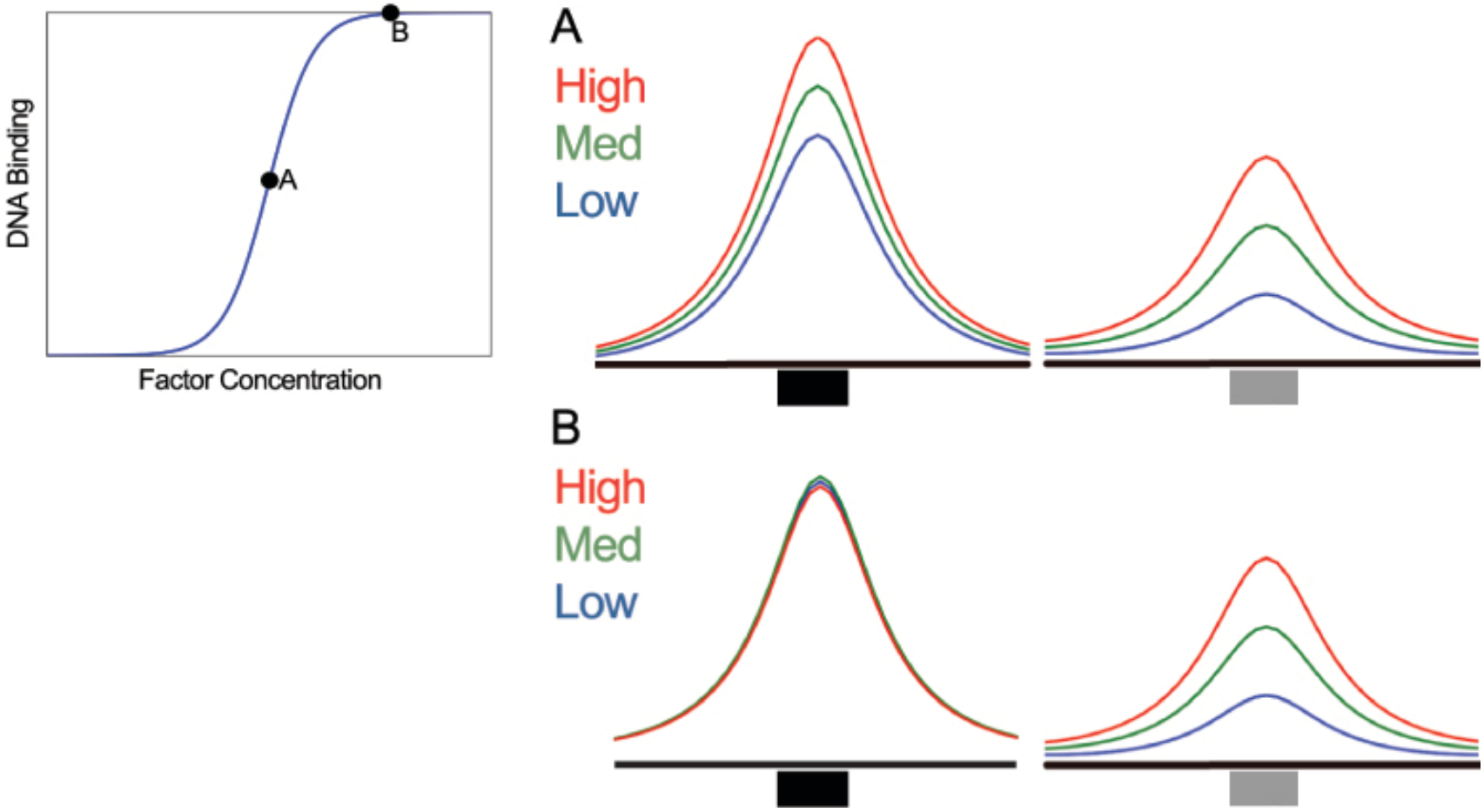

Figure 1: Two models for CRM response to increasing transcription factor. A) If *in vivo* concentration is near the kD for DNA binding, varying concentration of factor should effect changes in binding at both high (black bar) and low (grey bar) affinity bound regions. B) If *in vivo* concentration saturates the DNA binding curve, varying concentration should have no effect binding, though this may be mediated by the strength of the binding site.

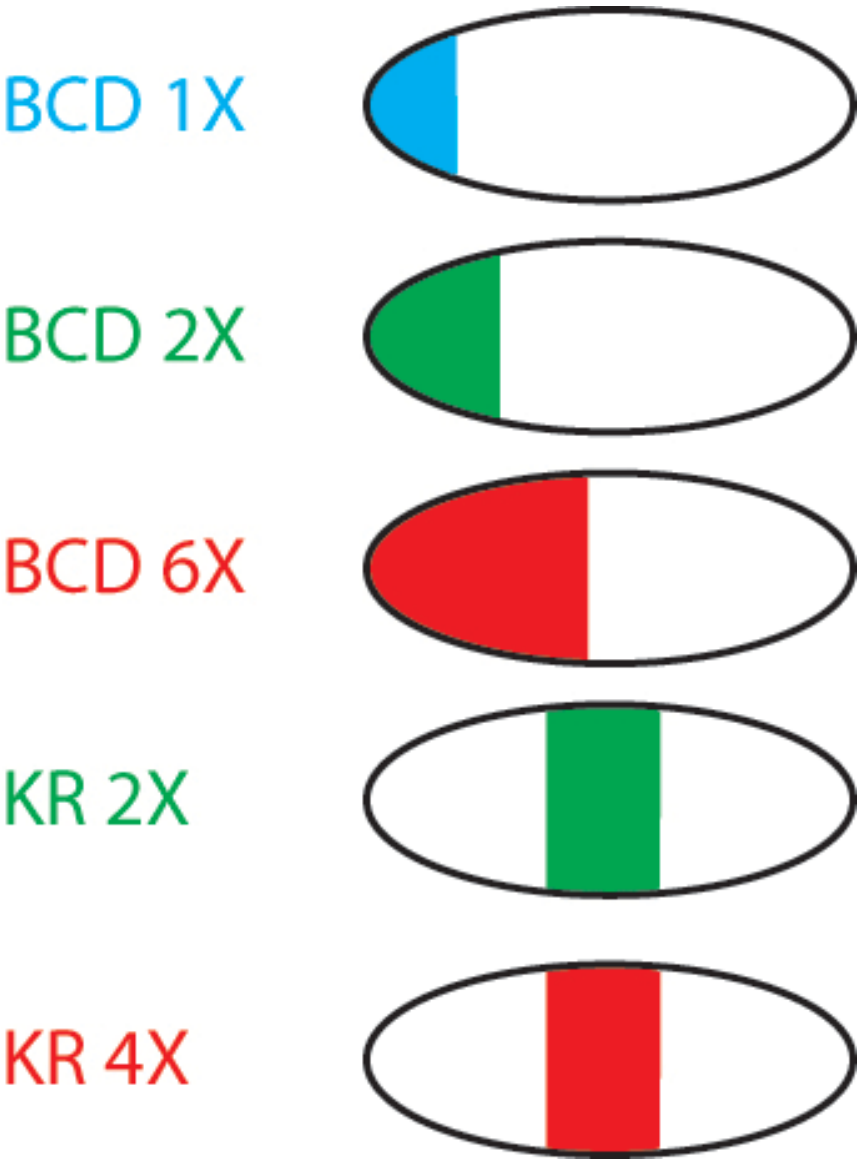

Figure 2: Mutant flies used in the dosage series for *bicoid* and *Kruppel* flies. FPKM for *bicoid* flies at Stage 4B were 34.0, 57.1, and 144.5 for 1X, 2X, and 6X flies. Protein levels agree with quantification by Driever and Nusslein-Volhard. *Kruppel* FPKMs were 68.9 and 198.6 for 2X and 4X flies. Both BCD and KR protein levels are awaiting quantification via the BDTNP microscopy pipeline.

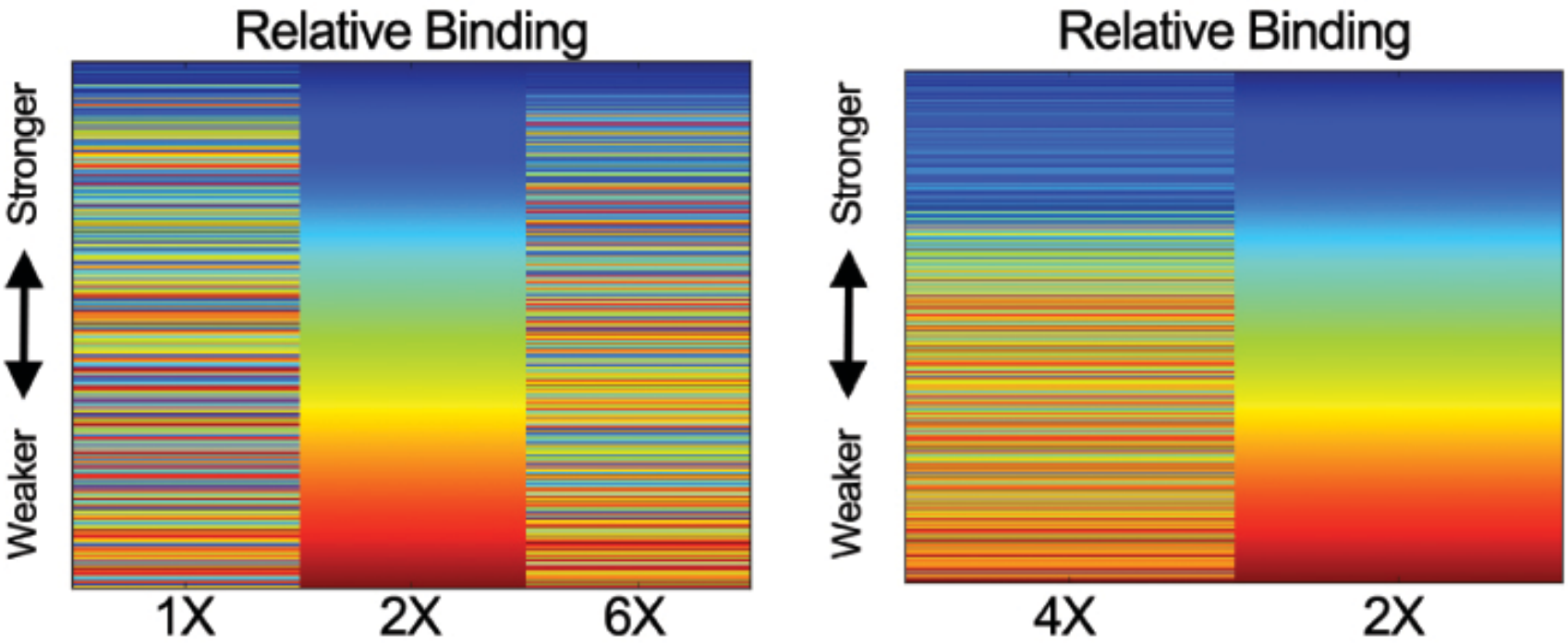

Figure 3: Binding rank changes with gene dosage for the 1% FDR bound regions. A) BCD 1X disruption is greater than 6X disruption in 392 BCD bound regions, but in both cases the top of the rank list is consistent. B) KR 4X dosage disrupts ranked binding throughout 1770 bound regions, but this disruption is less than the disruption by either BCD mutant.

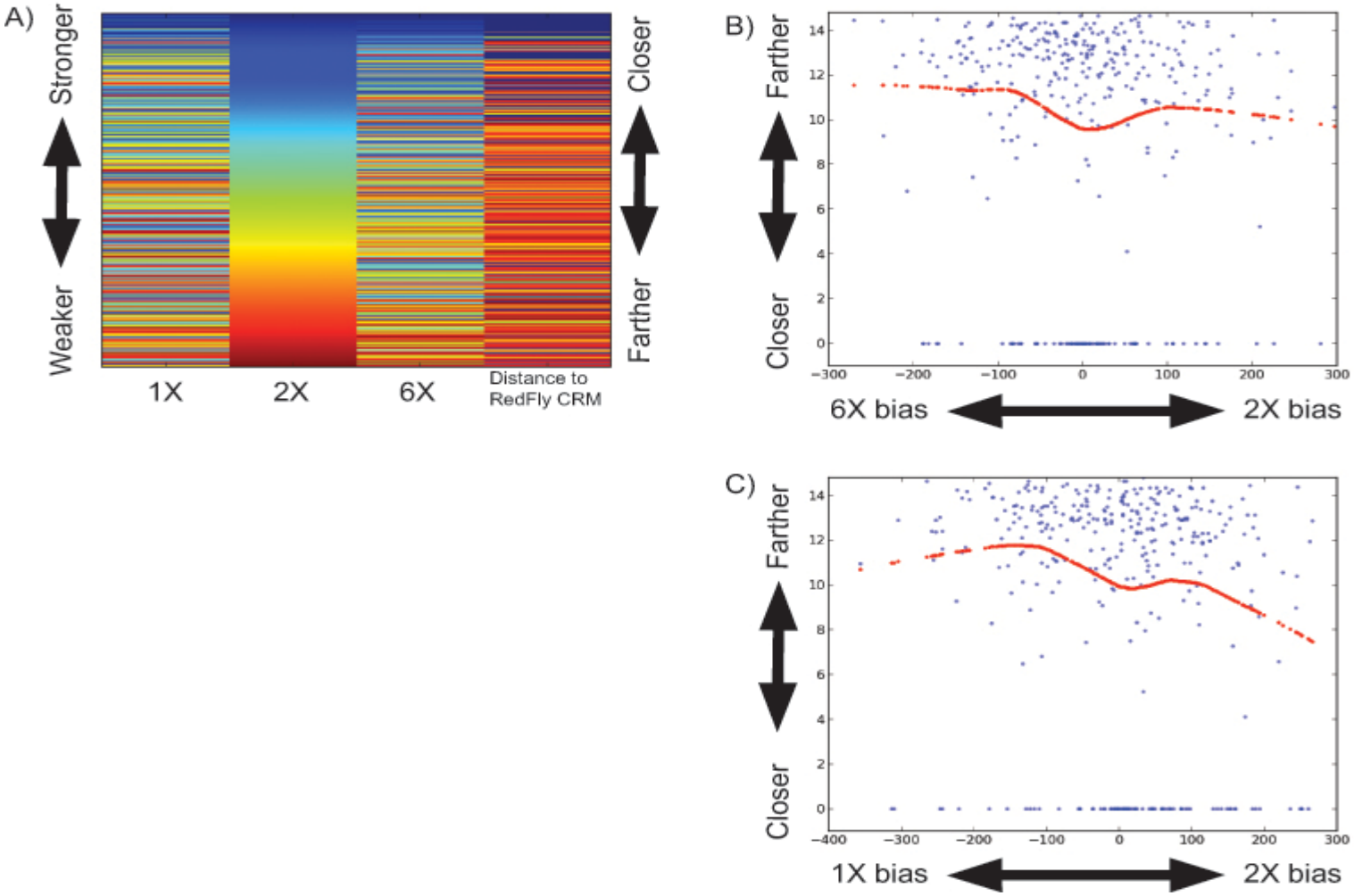

Figure 4: Highly bound regions are highly consistent and proximal to functionally annotated CRMs. A) Regions with stronger binding are much closer to RedFly regions. B&C) Overlapped Redfly regions are enriched in the high-consistency regions (array a data points at zero on the y-axis). Regions biased toward higher binding in the 6X and 1X sample are slightly farther from RedFly regions.

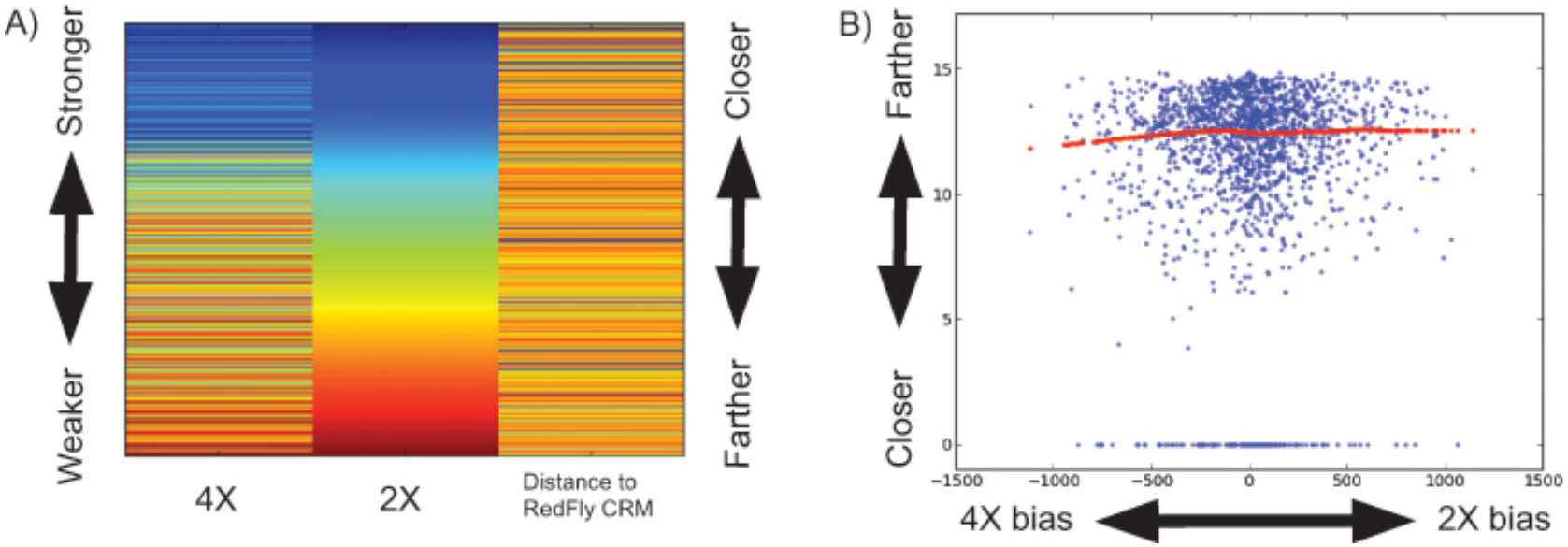

Figure 5: A) Unlike BCD regions, KR bound regions are not proximal to functionally annotated CRMs. B) Regions biased in the 4X or 2X data are unbiased with respect to RedFly proximity.

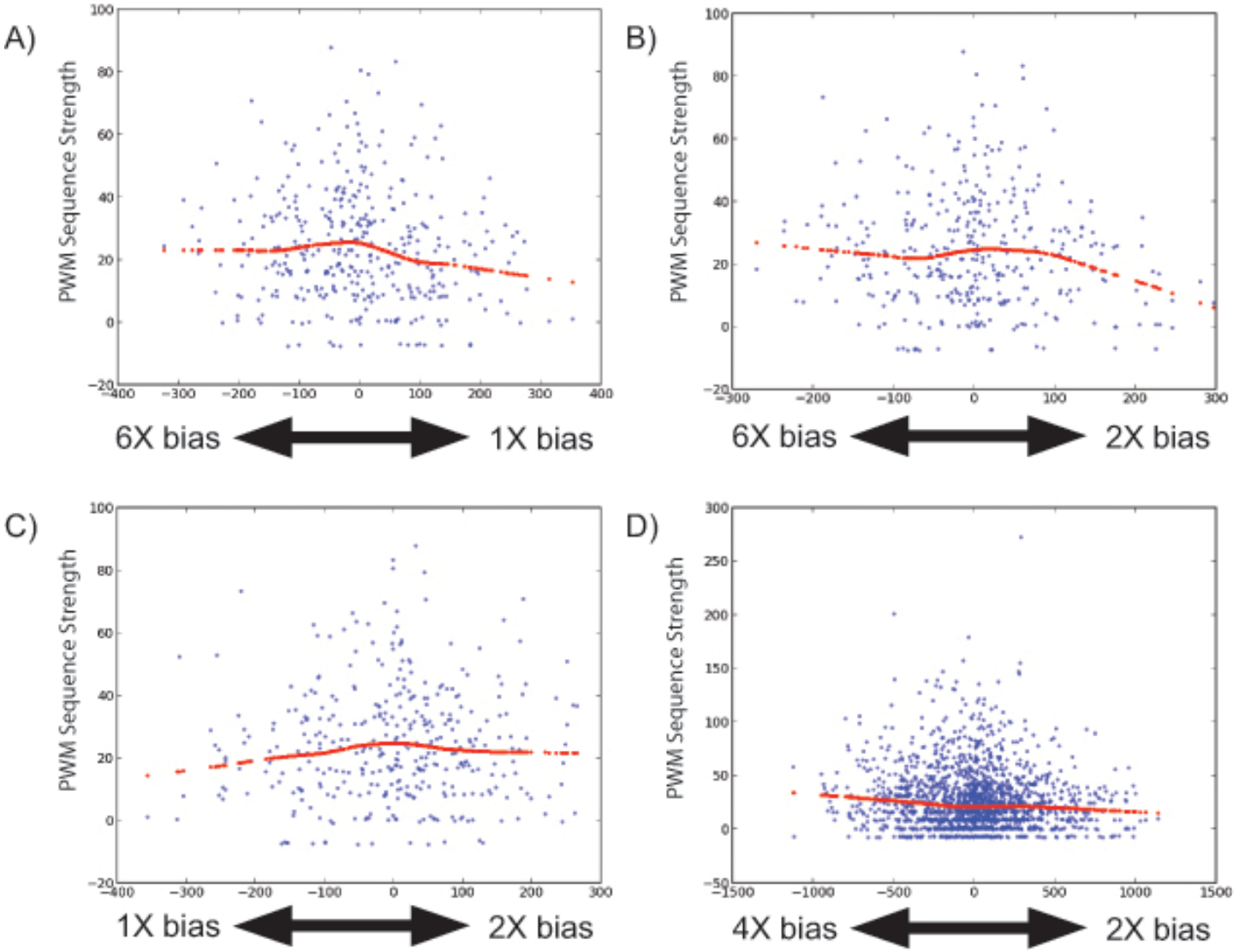

Figure 6: Sequence Affinity of rank-biased regions. A,B) BCD 6X biased regions have higher primary sequence affinity than 1X ror 2X based regions, but high-consistency regions have the highest sequence affinity of all. C) Symmetrically, 1X biased regions have lower affinity that 2X sequences, though again high-consistency regions have the highest BCD sequence affinity. D) Like BCD regions, KR 4X biased regions have higher affinity than 2X biased regions. Unlike BCD, the high-consistency regions do not have higher, but intermediate affinity for KR.

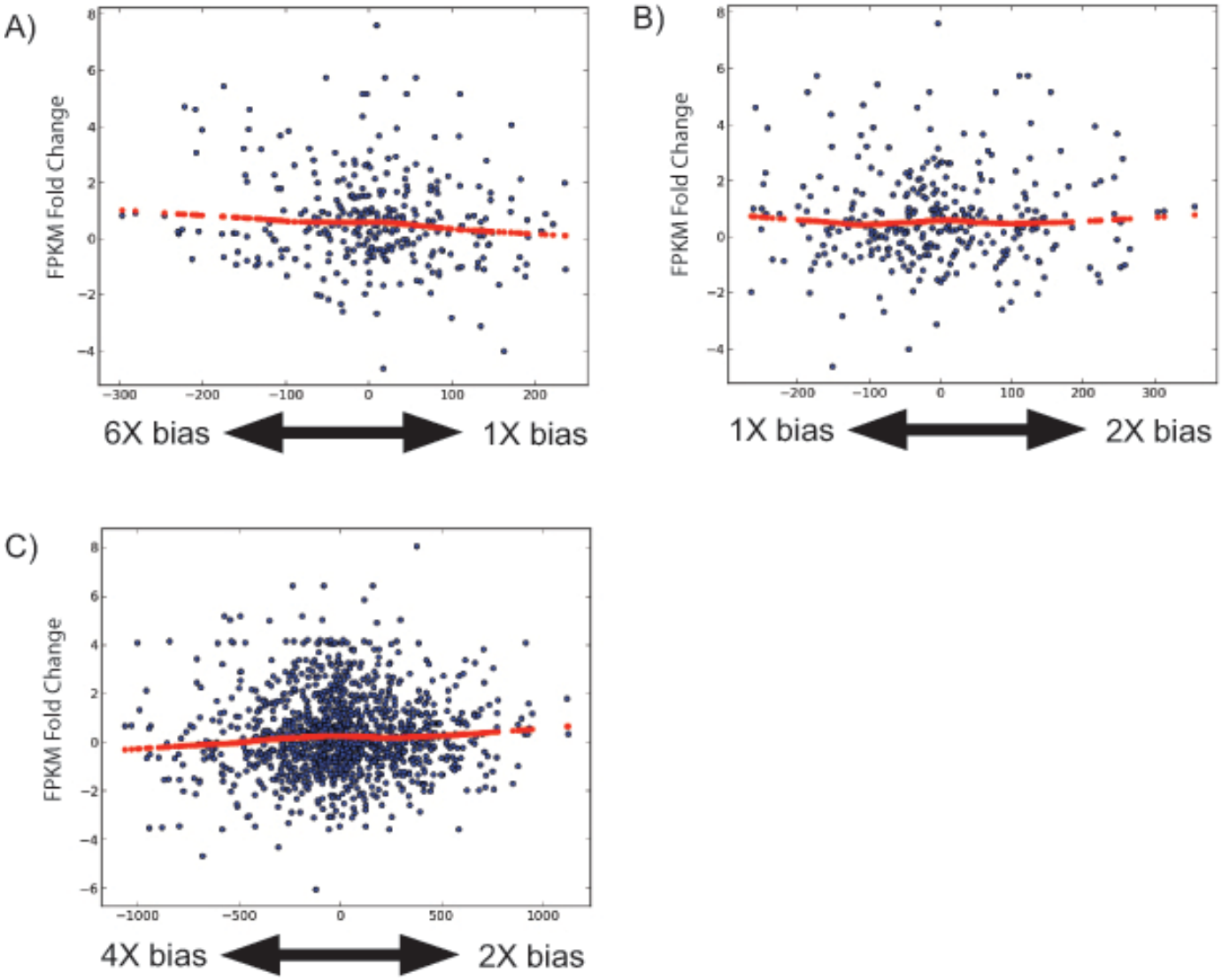

Figure 7: Changes in binding correspond to changes in expression of nearby genes. A) Genes near BCD 6X biased regions are expressed more highly than 1X biased genes. B) BCD 1X biased regions appear little or no effect on nearby gene expression. C) Genes near KR 4X biased regions are expressed less than high-consistency regions or genes near 2X biased regions.

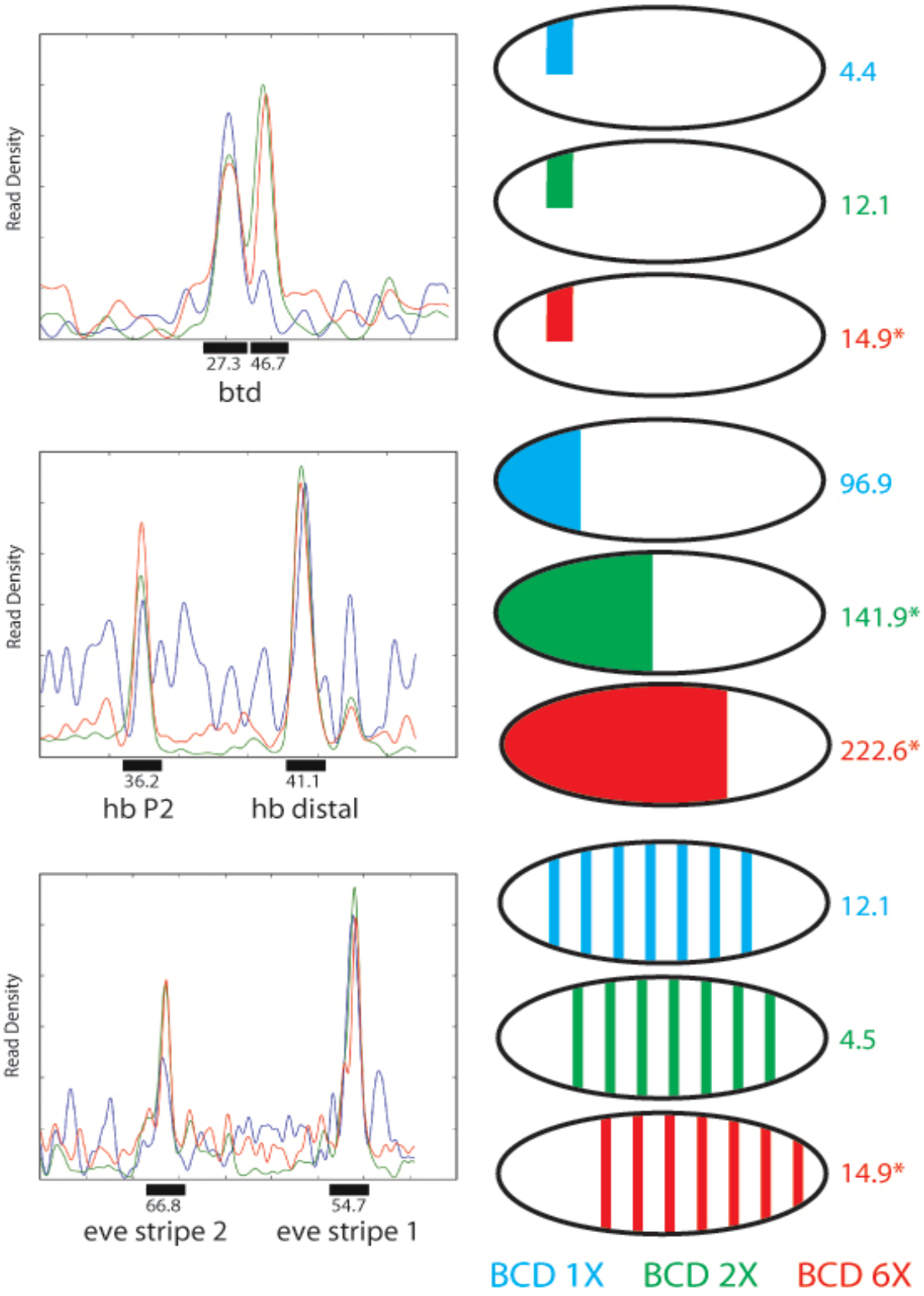

Figure 8: Changes in BCD binding at Functional Loci. A) Two binding site clusters at the *btd* locus differ in dosage sensitivity. *btd* expression domain is not dosage sensitive, but there is a minor increase in expression at 6X. B) *hb* P2 binding is dosage sensitive disproportionately to the distal enhancer. *hb* expression is very dosage sensitive, and FPKM reflects this. C) *eve* stripe 2 enhancer is dosage sensitive, but stripe 1 is not. *eve* pattern is dosage sensitive, but shifts rather than expanding

Chapter 4

Two Vignettes of Collaborations Studying Transcriptional Regulation After Formation of the RNA Polymerase II Open Complex

Transcriptional regulation continues after transcription begins. In this chapter, I briefly describe as vignettes two collaborative efforts studying transcriptional regulation after the stable recruitment of RNA Polymerase II to an actively transcribed gene. The first vignette concerns the yeast AAA-ATPase Yta7, a chromatin remodeling protein[208-211]. The second concerns the Super Elongation Complex (SEC)[107-110,212], a complex of proteins present at the transcription start site of many actively transcribed genes in human cells.

Nucleosome Reorganization in yta7 Mutant Yeasts

Yta7 protein is localized to highly transcribed loci where it facilitates transcription of the locus by decreasing nucleosome density and promoting transcript elongation[208-211]. Yta7 can directly bind histone H3, and under-expression of histones H3 and H4 attenuates the effect of the yta7 mutant. This evidence and the observation that yta7 mutants exhibit higher nucleosome density suggested that Yta7 was regulating the presence and density of nucleosomes at highly transcribed genes in the genome.

My collaborators MNase digested chromatin from yta7 over-expressors, yta7-delta mutants, and wild-type yeast. Illumina libraries were constructed with the resulting DNA fragments and the sample sequenced. The resulting reads were mapped to the reference *S. cerevisae* genome with *bwa-64*. The read density per sample was normalized to a total of 10,000,000 reads, and these densities were used to identify the genomic positions of nucleosomes with a combination of previously published[213] and custom software. The previous conclusions derived from the study of a handful of induced loci were supported by the genomic characterization of the yta7 mutant strains. Nucleosome density and spacing does not vary at the -1 position, but density shifts as expected with regard to wild-type data (Figure 1). Additionally, Yta7 is bound pervasively, but not constitutively, at highly expressed genes, where this effect is exacerbated (not shown).

Members of the AFF4 Scaffold-associated SEC are Differentially Localized and Facilitate High Transcriptional Output

RNA Polymerase II is poised at many genes that are not actively transcribed[101-103,214,215]. At these loci, some transcription occurs, but the nascent transcript cannot persistently nor efficiently elongate. Unpausing of this activated polymerase is coordinate with phosphorylation of the Ser2 residue of the C-terminal tail domain of the polymerase[103]. The P-Tefb complex includes cyclin T1 (CycT1) and the cyclin-dependent kinase CDK9[104]. This complex is responsible for the phosphorylation events that precipitate unpausing and efficient elongation of the transcript. Members of the AFF family of protein scaffolds organize P-Tefb with the other members of the SEC complex at the transcription start sites of highly induced genes[107]. It has been observed that at the mis-induced hox genes of MLL-fusion leukocyte leukemia culture cells, the AFF scaffolds are present throughout the transcribed ORF[216]. The MLL-fusion genes identified by their association with MLL leukemias are enriched for members of the SEC complex. From these observations,

it has been suggested that the SEC has a role in gene expression of developmental genes that when misexpressed contribute to cancer, and also that the SEC is genomically localized throughout these highly expressed loci.

My collaborators have collected high-throughput sequencing data in human HeLa cell culture measuring the genomic localization by immunoprecipitation of the biochemically verified SEC complex members AF9, ELL2, and ENL, along with the scaffolding proteins AFF1/4, the P-TefB members CycT1 and CDK9, the P-Tefb-associated bromo-domain protein Brd4, and general and phospho-specific RNA Polymerase II complexes. In each case, these proteins bind at thousands of loci in the genome, and are deeply enriched for actively transcribed regions (Figure 2, 4). However, these proteins are not bound coordinately in every case, but rather some combinatorially near different transcription start sites (Figure 2). All members are present at the TSS of actively transcribed genes, but only some at genes with paused polymerase. Notably, the SEC does not appear to extend deep into the locus of actively transcribed genes. Rather, the complex is typically bound well into the first exon of actively transcribed genes, but only ENL and AF9 show any evidence of binding deeper into loci. This is in contrast to the model extrapolated from the hox loci. However, the hox loci are uncharacteristically short human genes, averaging only a few kb in length, and containing no introns. Thus, it appears that the SEC typically elongates 2-3kb into the first exon of actively transcribed loci, and in the case of hox genes, this happens to be the entire locus (Figure 3).

To assess the effect of the SEC components on transcriptional output, my collaborators collected high-throughput sequencing data measuring the transcriptomes of HeLa cell cultures after treatment with RNAi constructs for combinations of CDK9, AFF1 and AFF4, and ENL and AF9. These represent one dataset with RNAi targeted to an elongation pre-requisite (CDK9), the scaffold proteins themselves (AFF1/4), and members that may extend more deeply into actively transcribed loci. These datasets clearly show effective knockdown of their intended targets, and the suppression of gene expression of thousands of loci (Table 2) compared to expression in untreated cells. When these data are considered with the binding data, it is clear that TSS bound in untreated cells by are on average expressed far higher than average loci, consistent with the role of the SEC as a transcriptional upregulator (Figure 4). However, the effect on gene expression is not the same for each component. Indeed, the components that extend the most deeply into the first exon, AF9, ELL2, and ENL, along with CDK9, are associated with the highest gene expression, followed by Brd4, CycT1, and the two scaffold proteins themselves. Though this intriguing observation begs the question of stronger transcriptional promotion by more deeply elongating factors, but this conclusion should be tempered. The efficiency of the chromatin immunoprecipitation varies for each of these factors, and it is possible that the appearance of deeper extension into actively transcribed regions may be an artifact of poorer efficiency in the ChIP. This would suggest that the entire SEC does indeed extend into the locus, but only the first 2-3kb. Either explanation is a possibility that is difficult to distinguish in the

case of chromatin associated proteins without sequence-specific recognition site to orthogonally justify the differential elongation model.

gene	control	RNAi CDK9	RNAi AFF1-4	RNAi ENL-AF9
CDK9	5.75	1.26	4.57	3.69
AFF1	3.84	2.60	2.46	5.61
AFF4	12.28	10.37	3.31	14.32
ENL	17.61	14.02	12.18	3.03
AF9	17.46	11.95	13.75	3.30

Table 8A: Knockdown efficiency was high for each SEC member.

Sample	> 2 fold decrease
Control (no treatment)	n/a
RNAi CDK9	2302
RNAi AFF1/4	1605
RNAi ENL/AF9	2001

Table 1B: Expression of thousands of targets was suppressed with RNAi treatment.

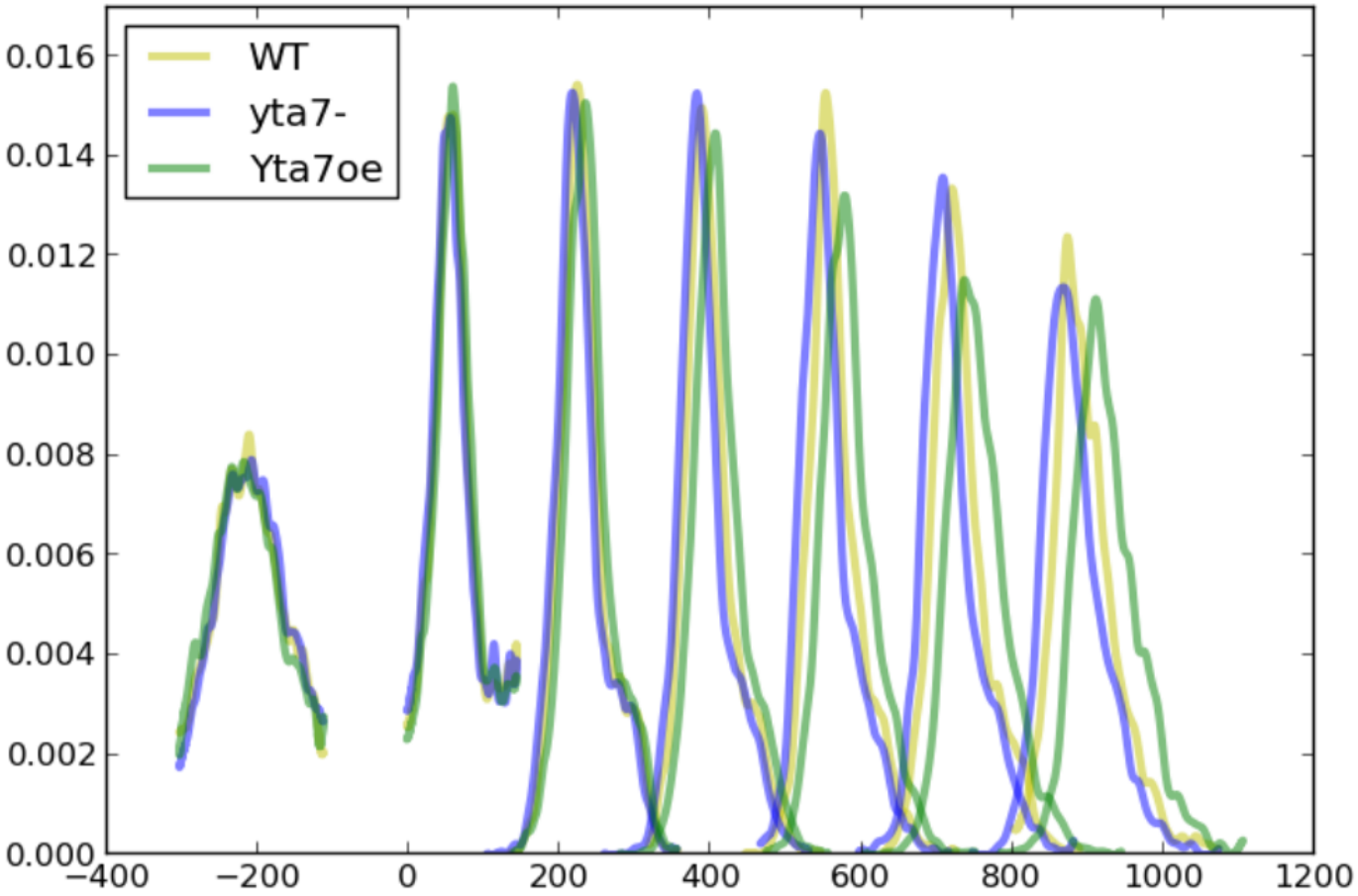

Figure 1: Normalized and smoothed nucleosome density across the meta-TSS (NFR omitted). While the -1 is similarly positioned across loci, YTA7 levels shift the nucleosomes after the TSS relative the wild-type.

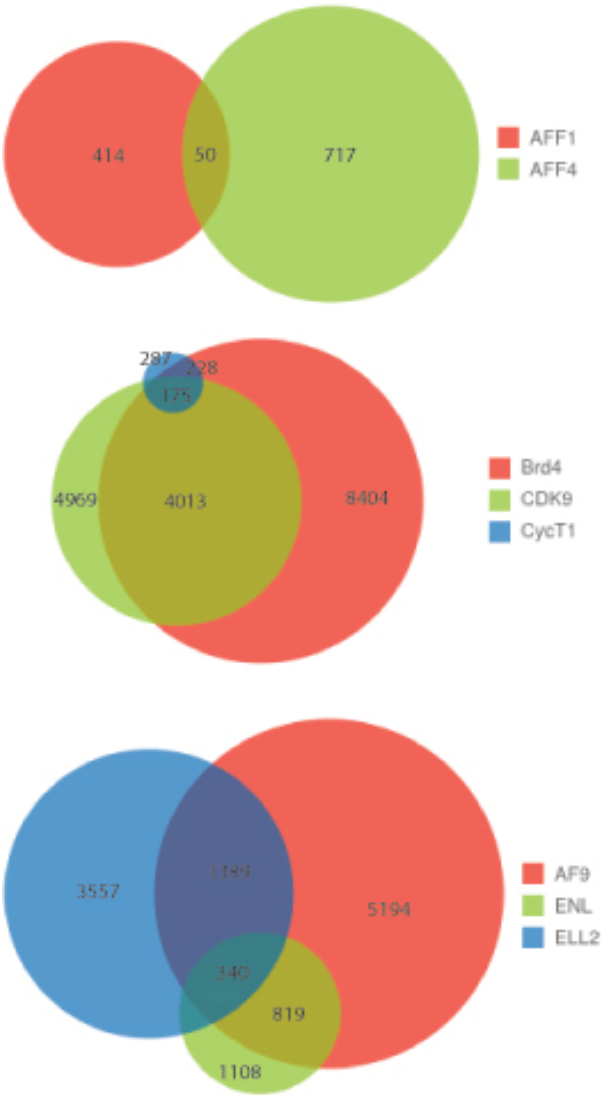

Figure 2: Venn diagrams depicting overlap in three functional groups of SEC proteins and their targets. A) AFF1 and AFF4 scaffolds appear regulate largely different targets. B) CDK9 is present at most Brd4 bound sites, but CycT1 is detectable at an order of magnitude fewer sites. C) AF9, ELL, and ENL overlap commonly, but ENL binding is most often accompanied by AF9 binding.

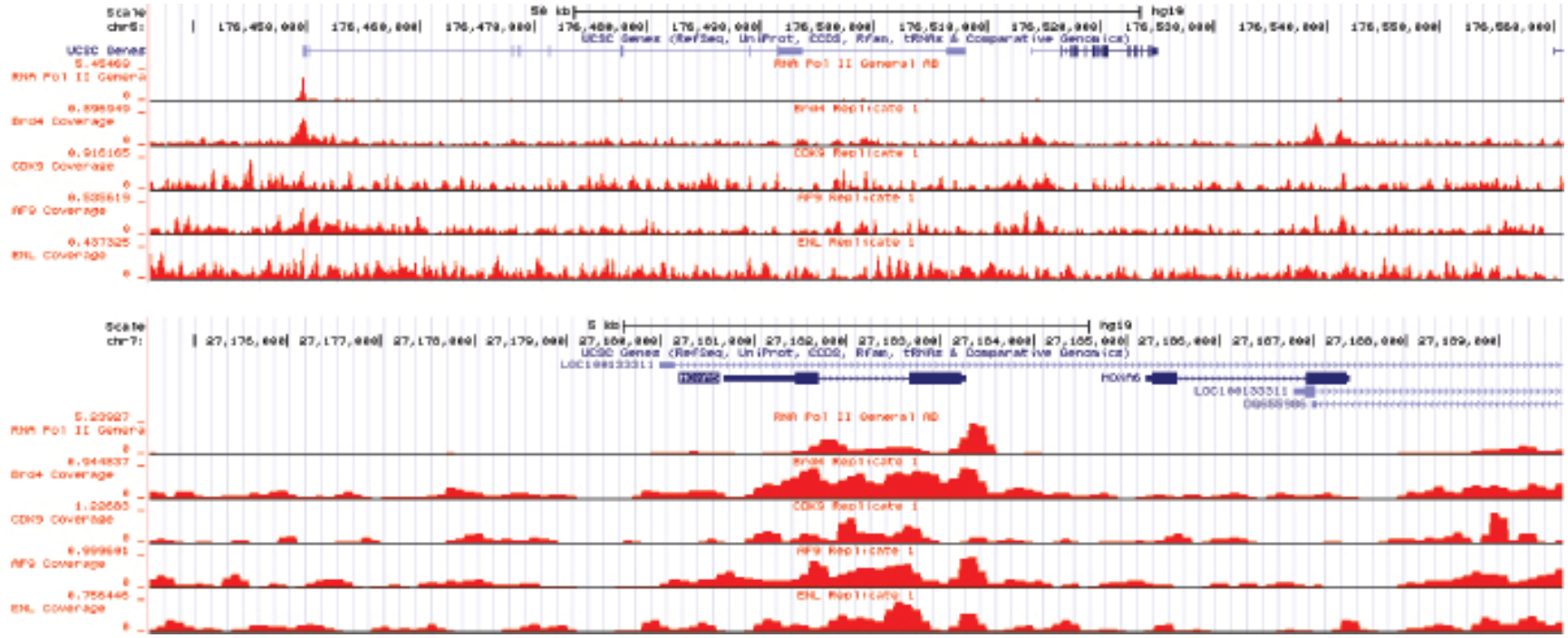

Figure 3: The SEC rarely extends deep into highly transcribed loci. Top) ZNF346, a highly expressed gene of 58.4kb is bound within the first few kb of the locus by RNA Pol II and the SEC. Bottom) HOXA5, a highly expressed locus of 2.6kb is bound across the entire locus, which also happens to be the first few kb of the locus.

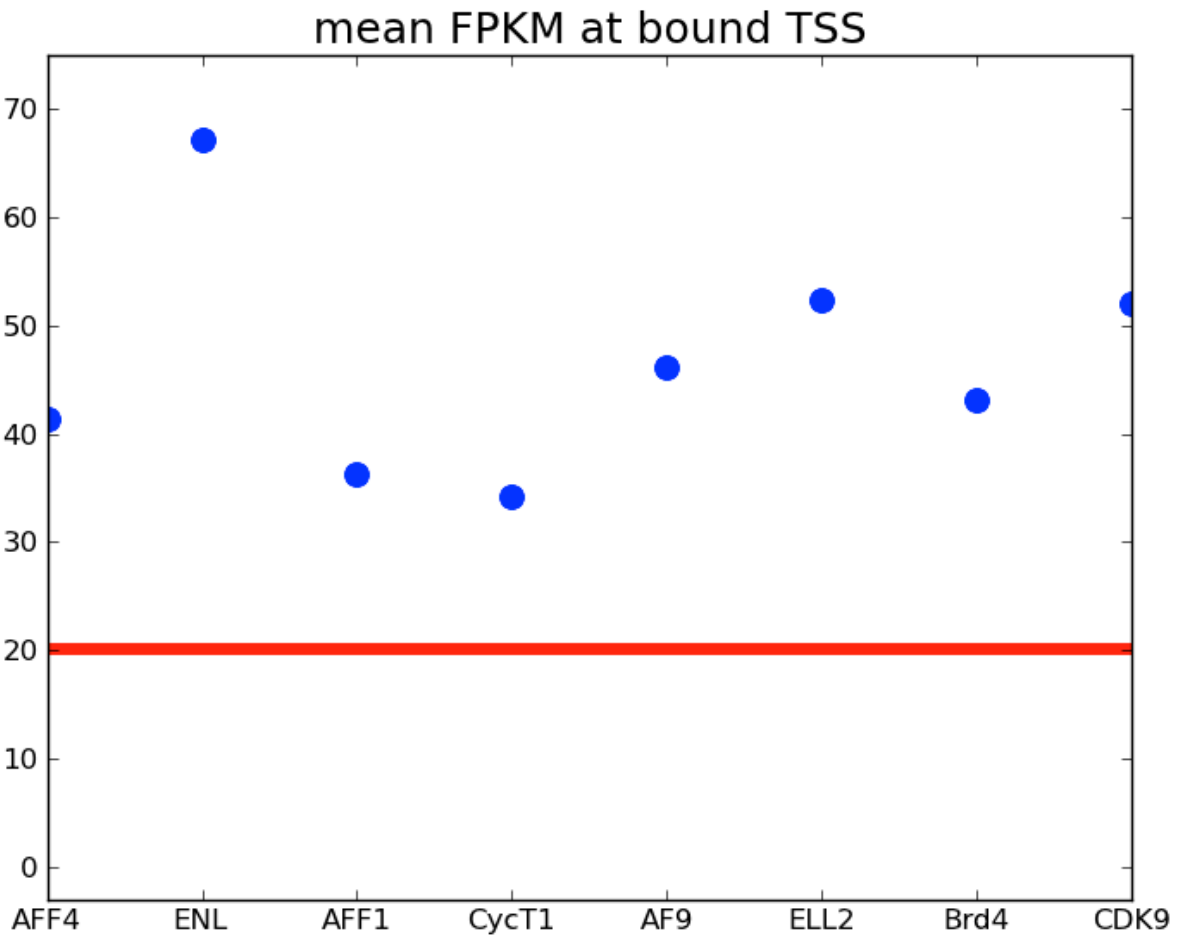

Figure 4: Different SEC members are associated with variably high transcript abundance. The red line near 20 FPKM represents the mean FPKM, and SEC bound loci are all expressed at high mean FPKM. ENL targets are especially highly expressed.

References

1. Judson, H. F. *The eighth day of creation*. (Simon & Schuster, 1980).
2. Nirenberg, M. *et al.* RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc Natl Acad Sci USA* **53**, 1161–1168 (1965).
3. GARDNER, R. S. *et al.* Synthetic polynucleotides and the amino acid code. VII. *Proc Natl Acad Sci USA* **48**, 2087–2094 (1962).
4. CRICK, F. H., BARNETT, L., Brenner, S. & WATTS-TOBIN, R. J. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
5. Turing, A. M. The chemical basis of morphogenesis. *Phil Trans Roy Soc* (1952).
6. C. elegans Sequencing Consortium Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
7. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546–563–7 (1996).
8. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–2195 (2000).
9. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
10. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
11. Carroll, S. B. *Endless Forms Most Beautiful*. (W. W. Norton, 2006).
12. Davidson, E. H. *Genomic Regulatory Systems*. (Academic Press, 2001).
13. Carroll, S. B. Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577–580 (2000).
14. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
15. Rubin, G. M. Comparative Genomics of the Eukaryotes. *Science* **287**, 2204–2215 (2000).
16. Postlethwait, J. H. Zebrafish Comparative Genomics and the Origins of Vertebrate Chromosomes. *Genome Res* **10**, 1890–1902 (2000).
17. Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T. & Brent, M. R. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS biology* (2003).
18. Stark, A. *et al.* Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
19. Drosophila 12 Genomes Consortium *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203–218 (2007).
20. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034–1050 (2005).
21. Dubchak, I. Active Conservation of Noncoding Sequences Revealed by Three-Way Species Comparisons. *Genome Res* **10**, 1304–1306 (2000).
22. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
23. Ronald, J., Brem, R. B., Whittle, J. & Kruglyak, L. Local regulatory variation in

Saccharomyces cerevisiae. *PLoS Genet* **1**, e25 (2005).
24. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. *Nature* **430**, 85–88 (2004).
25. Rockman, M. V. *et al.* Ancient and recent positive selection transformed opioid cis-regulation in humans. *PLoS Biol* **3**, e387 (2005).
26. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
27. Veyrieras, J.-B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* **4**, e1000214 (2008).
28. Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20**, 1377–1419 (2003).
29. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**, 206–216 (2007).
30. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95–108 (2005).
31. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
32. Lieb, J. D., Liu, X., Botstein, D. & Brown, P. O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* (2001).
33. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
34. Sabo, P. J. *et al.* Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci USA* **101**, 16837–16842 (2004).
35. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Meth* **3**, 17–21 (2006).
36. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc* **2**, 988–1002 (2007).
37. JACOB, F. & MONOD, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318–356 (1961).
38. Nurse, P. A Long Twentieth Century of Review the Cell Cycle and Beyond. *Cell* (2000).
39. Johnson, A. D. & Herskowitz, I. A repressor (MAT alpha 2 Product) and its operator control expression of a set of cell type specific genes in yeast. *Cell* **42**, 237–247 (1985).
40. McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A. & Gehring, W. J. A homologous protein-coding sequence in drosophila homeotic genes and its conservation in other metazoans. *Cell* **37**, 403–408 (1984).
41. Lewis, E. B. A gene complex controlling segmentation in Drosophila. *Nature* **276**, 565–570 (1978).
42. Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795–801 (1980).
43. Scott, M. P. *et al.* The molecular organization of the Antennapedia locus of Drosophila. *Cell* **35**, 763–776 (1983).
44. Driever, W. & Nüsslein-Volhard, C. A gradient of bicoid protein in Drosophila

embryos. *Cell* **54**, 83–93 (1988).
45. Blackwood, E. M. Going the Distance: A Current View of Enhancer Action. *Science* **281**, 60–63 (1998).
46. Kornberg, R. D. Eukaryotic transcriptional control. *Trends Biochem Sci* **24**, M46–M49 (1999).
47. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nat Genet* **39**, 1507–1511 (2007).
48. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39**, 1512–1516 (2007).
49. Steinmetz, E. J. *et al.* Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Mol Cell* **24**, 735–746 (2006).
50. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **8**, 424–436 (2007).
51. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
52. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–373 (2011).
53. Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M. D. & Hughes, T. R. Genome-Wide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors. *… and cellular biology* (2004).
54. Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci USA* **99**, 5860–5865 (2002).
55. Munchel, S. E., Shultzaberger, R. K., Takizawa, N. & Weis, K. Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Mol Biol Cell* **22**, 2787–2795 (2011).
56. Mata, J., Marguerat, S. & Bähler, J. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci* **30**, 506–514 (2005).
57. Bachmair, A. In vivo half-life of a protein is a function of its. *Science* (1986).
58. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281–1295 (1987).
59. Wilusz, C. J. & Wilusz, J. Bringing the role of mRNA decay in the control of gene expression into focus. *Trends Genet* **20**, 491–497 (2004).
60. Belle, A., Tanay, A., Bitincka, L., Shamir, R. & O'Shea, E. K. Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci USA* **103**, 13004–13009 (2006).
61. Eden, E. *et al.* Proteome half-life dynamics in living human cells. *Science* **331**, 764–768 (2011).
62. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707–719 (2007).
63. Santisteban, M. S., Kalashnikova, T. & Smith, M. M. Histone H2A.Z Regulates Transcription and Is Partially Redundant with Nucleosome Remodeling Complexes. *Cell* **103**, 411–422 (2000).

64. Berger, S. L. Histone modifications in transcriptional regulation. *Curr Opin Genet Dev* (2002).
65. Cairns, B. R. The logic of chromatin architecture and remodelling at promoters. *Nature* **461**, 193–198 (2009).
66. Martens, J. A. & Winston, F. Recent advances in understanding chromatin remodeling by Swi/Snf complexes. *Curr Opin Genet Dev* (2003).
67. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
68. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2008).
69. Peckham, H. E. *et al.* Nucleosome positioning signals in genomic DNA. *Genome Res* **17**, 1170–1177 (2007).
70. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* **18**, 1073–1083 (2008).
71. Valouev, A. *et al.* A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *… research* (2008).
72. Tjian, R. The binding site on SV40 DNA for a T antigen-related protein. *Cell* **13**, 165–179 (1978).
73. Struhl, K. Deletion mapping a eukaryotic promoter. *Proc Natl Acad Sci USA* **78**, 4461–4465 (1981).
74. McKnight, S. L. & Kingsbury, R. Transcriptional control signals of a eukaryotic protein-coding gene. *Science* **217**, 316–324 (1982).
75. Kadonaga, J. T. Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors. *Cell* (2004).
76. O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* **254**, 539–544 (1991).
77. Hai, T. & Curran, T. Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *Proc Natl Acad Sci USA* **88**, 3720–3724 (1991).
78. Pabo, C. O. & Sauer, R. T. Transcription factors: structural families and principles of DNA recognition. *Annu Rev Biochem* (1992).
79. Brent, R. & Ptashne, M. A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* **43**, 729–736 (1985).
80. Ptashne, M. Gene regulation by proteins acting nearby and at a distance. *Nature* **322**, 697–701 (1986).
81. Ptashne, M. How eukaryotic transcriptional activators work. *Nature* **335**, 683–689 (1988).
82. Ptashne, M. Principles of a switch. *Nat. Chem. Biol.* **7**, 484–487 (2011).
83. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
84. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J Mol Biol* **188**, 415–431 (1986).
85. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097–6100 (1990).

86. Biggin, M. D. Animal transcription networks as highly connected, quantitative continua. *Dev Cell* **21**, 611–626 (2011).
87. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* **43**, 264–268 (2011).
88. Voss, T. C. *et al.* Dynamic Exchange at Regulatory Elements during Chromatin Remodeling Underlies Assisted Loading Mechanism. *Cell* (2011).doi:10.1016/j.cell.2011.07.006
89. Biddie, S. C. *et al.* Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell* **43**, 145–155 (2011).
90. Slattery, M. *et al.* Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. *Cell* **147**, 1270–1282 (2011).
91. Joshi, R. *et al.* Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131**, 530–543 (2007).
92. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* **79**, 233–269 (2010).
93. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248–1253 (2009).
94. Mann, R. S., Lelli, K. M. & Joshi, R. Hox specificity unique roles for cofactors and collaborators. *Current Topics in Developmental Biology* **88**, 63–101 (2009).
95. Blobel, G. A., Nakajima, T., Eckner, R., Montminy, M. & Orkin, S. H. CREB-binding protein cooperates with transcription factor GATA-1 and is required for erythroid differentiation. *Proceedings of the …* (1998).
96. Ryu, S., Zhou, S., Ladurner, A. G. & Tjian, R. The transcriptional cofactor complex CRSP is required for activity of the enhancer-binding protein Sp1. *Nature* **397**, 446–450 (1999).
97. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
98. Nibu, Y. & Levine, M. S. CtBP-dependent activities of the short-range Giant repressor in the Drosophila embryo. *Proc Natl Acad Sci USA* **98**, 6204–6208 (2001).
99. Nibu, Y. *et al.* dCtBP mediates transcriptional repression by Knirps, Krüppel and Snail in the Drosophila embryo. *EMBO J* **17**, 7009–7020 (1998).
100. Stanojević, D., Small, S. & Levine, M. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science* **254**, 1385–1387 (1991).
101. Lee, H., Kraus, K. W., Wolfner, M. F. & Lis, J. T. DNA sequence requirements for generating paused polymerase at the start of hsp70. *Genes Dev* **6**, 284–295 (1992).
102. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186–192 (2009).
103. O'Brien, T., Hardin, S., Greenleaf, A. & Lis, J. T. Phosphorylation of RNA polymerase II C-terminal domain and transcriptional elongation. *Nature* **370**, 75–77 (1994).
104. Price, D. H. P-TEFb, a Cyclin-Dependent Kinase Controlling Elongation by

RNA Polymerase II. *Mol Cell Biol* (2000).
105. Yang, Z. *et al.* Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol Cell* **19**, 535–545 (2005).
106. Jang, M. K. *et al.* The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell* **19**, 523–534 (2005).
107. He, N. *et al.* Human Polymerase-Associated Factor complex (PAFc) connects the Super Elongation Complex (SEC) to RNA polymerase II on chromatin. *Proc Natl Acad Sci USA* **108**, E636–45 (2011).
108. Luo, Z. *et al.* The SEC family of RNA Polymerase II elongation factors: gene target specificity and transcriptional output. *Mol Cell Biol* (2012).doi:10.1128/MCB.00182-12
109. Lin, C. *et al.* AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. *Mol Cell* **37**, 429–437 (2010).
110. Takahashi, H. *et al.* Human Mediator Subunit MED26 Functions as a Docking Site for Transcription Elongation Factors. *Cell* **146**, 92–104 (2011).
111. Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A. & Crossland, J. P. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* **313**, 101–104 (2006).
112. Manceau, M., Domingues, V. S., Mallarino, R. & Hoekstra, H. E. The developmental role of Agouti in color pattern evolution. *Science* **331**, 1062–1065 (2011).
113. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995–1016 (2007).
114. Liu, Z., Scannell, D. R., Eisen, M. B. & Tjian, R. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* **146**, 720–731 (2011).
115. Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**, 3157–3170 (1978).
116. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* (1990).
117. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
118. Joyce, G. F. Amplification, mutation and selection of catalytic RNA. *Gene* **82**, 83–87 (1989).
119. Meng, X., Brodsky, M. H. & Wolfe, S. A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat Biotechnol* **23**, 988–994 (2005).
120. Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* **36**, 1331–1339 (2004).
121. Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).

122. Li, X.-Y. *et al.* Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* **6**, e27 (2008).
123. Pepke, S., Wold, B. & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat Meth* **6**, S22–32 (2009).
124. Bateman, J. R., Lee, A. M. & Wu, C.-T. Site-specific transformation of Drosophila via phiC31 integrase-mediated cassette exchange. *Genetics* **173**, 769–777 (2006).
125. Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
126. Bergman, C. M., Carlson, J. W. & Celniker, S. E. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. (2005).
127. Kaplan, T. *et al.* Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet* **7**, e1001290 (2011).
128. He, X., Samee, M. A. H., Blatti, C. & Sinha, S. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol* **6**, (2010).
129. Breiman, L. *Classification and regression trees*. (1984).
130. Jaeger, J. *et al.* Dynamic control of positional information in the early Drosophila embryo. *Nature* **430**, 368–371 (2004).
131. Perkins, T. J., Jaeger, J., Reinitz, J. & Glass, L. Reverse engineering the gap gene network of Drosophila melanogaster. *PLoS Comput Biol* **2**, e51 (2006).
132. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**, 480–485 (2011).
133. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104 (2004).
134. Levine, M. & Manley, J. L. Transcriptional repression of eukaryotic promoters. *Cell* **59**, 405–408 (1989).
135. Small, S., Kraut, R., Hoey, T., Warrior, R. & Levine, M. Transcriptional regulation of a pair-rule stripe in Drosophila. *Genes Dev* **5**, 827–839 (1991).
136. Gray, S., Szymanski, P. & Levine, M. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. *Genes Dev* **8**, 1829–1838 (1994).
137. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205–214 (1996).
138. Revyakin, A. *et al.* Transcription initiation by human RNA polymerase II visualized at single-molecule resolution. *Genes Dev* **26**, 1691–1702 (2012).
139. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**, 166–176 (2003).
140. Laughon, A. & Scott, M. P. Sequence of a Drosophila segmentation gene: protein structure homology with DNA-binding proteins. *Nature* **310**, 25–31

(1984).
141. Struhl, G., Johnston, P. & Lawrence, P. A. Control of Drosophila body pattern by the hunchback morphogen gradient. *Cell* **69**, 237–249 (1992).
142. Mann, R. S. & Hogness, D. S. Functional dissection of Ultrabithorax proteins in D. melanogaster. *Cell* **60**, 597–610 (1990).
143. Sauer, F. & Jäckle, H. Concentration-dependent transcriptional activation or repression by Krüppel from a single binding site. *Nature* **353**, 563–566 (1991).
144. Grimm, O., Coppey, M. & Wieschaus, E. Modelling the Bicoid gradient. *Development* **137**, 2253–2264 (2010).
145. Berman, B. P. *et al.* Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol* **5**, R61 (2004).
146. MacArthur, S. *et al.* Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol* **10**, R80 (2009).
147. Bradley, R. K. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol* **8**, e1000343 (2010).
148. Hülskamp, M. & Tautz, D. Gap genes and gradients--the logic behind the gaps. *Bioessays* **13**, 261–268 (1991).
149. Jäckle, H. *et al.* Transcriptional control by Drosophila gap genes. *J Cell Sci Suppl* **16**, 39–51 (1992).
150. Pankratz, M. J. & Jäckle, H. Making stripes in the Drosophila embryo. *Trends Genet* **6**, 287–292 (1990).
151. Mannervik, M. Target genes of homeodomain proteins. *Bioessays* **21**, 267–270 (1999).
152. Luengo Hendriks, C. L. *et al.* Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. *Genome Biol* **7**, R123 (2006).
153. Keränen, S. V. E. *et al.* Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution II: dynamics. *Genome Biol* **7**, R124 (2006).
154. Fowlkes, C. C. *et al.* A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell* **133**, 364–374 (2008).
155. Struhl, G., Struhl, K. & Macdonald, P. M. The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell* **57**, 1259–1273 (1989).
156. Wolpert, L. Positional information and the spatial pattern of cellular differentiation. *J Theor Biol* **25**, 1–47 (1969).
157. Eldon, E. D. & Pirrotta, V. Interactions of the Drosophila gap gene giant with maternal and zygotic pattern-forming genes. *Development* **111**, 367–378 (1991).
158. Ronchi, E., Treisman, J., Dostatni, N., Struhl, G. & Desplan, C. Down-regulation

of the Drosophila morphogen bicoid by the torso receptor-mediated signal transduction cascade. *Cell* **74**, 347–355 (1993).
159. Kraut, R. & Levine, M. Mutually repressive interactions between the gap genes giant and Krüppel define middle body regions of the Drosophila embryo. *Development* **111**, 611–621 (1991).
160. Papatsenko, D. & Levine, M. S. Dual regulation by the Hunchback gradient in the Drosophila embryo. *Proc Natl Acad Sci USA* **105**, 2901–2906 (2008).
161. Sauer, F., Hansen, S. K. & Tjian, R. DNA template and activator-coactivator requirements for transcriptional synergism by Drosophila bicoid. *Science* **270**, 1825–1828 (1995).
162. Margolis, J. S. *et al.* Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element. *Development* **121**, 3067–3077 (1995).
163. Hiromi, Y., Kuroiwa, A. & Gehring, W. J. Control elements of the Drosophila segmentation gene fushi tarazu. *Cell* **43**, 603–613 (1985).
164. Hiromi, Y. & Gehring, W. J. Regulation and function of the Drosophila segmentation gene fushi tarazu. *Cell* **50**, 963–974 (1987).
165. Ish-Horowicz, D., Pinchin, S. M., Ingham, P. W. & Gyurkovics, H. G. Autocatalytic ftz activation and metameric instability induced by ectopic ftz expression. *Cell* **57**, 223–232 (1989).
166. Fujioka, M., Emi-Sarker, Y., Yusibova, G. L., Goto, T. & Jaynes, J. B. Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development* **126**, 2527–2538 (1999).
167. Small, S., Blair, A. & Levine, M. Regulation of even-skipped stripe 2 in the Drosophila embryo. *EMBO J* **11**, 4047–4057 (1992).
168. Small, S., Blair, A. & Levine, M. Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. *Dev Biol* **175**, 314–324 (1996).
169. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R news* (2002).
170. Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. *Development* **130**, 6569–6575 (2003).
171. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94**, 890–898 (2005).
172. Therneau, T. M. & Atkinson, E. J. An introduction to recursive partitioning using the RPART routines. (1997).
173. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet* **4**, e1000106 (2008).
174. Fowlkes, C. C. *et al.* A conserved developmental patterning network produces quantitatively different output in multiple species of Drosophila. *PLoS Genet* **7**, e1002346 (2011).
175. Wittkopp, P. J. *et al.* Intraspecific polymorphism to interspecific divergence: genetics of pigmentation in Drosophila. *Science* **326**, 540–544 (2009).
176. Miller, C. T. *et al.* cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell* **131**, 1179–1189

(2007).
177. Ludwig, M. Z. *et al.* Functional evolution of a cis-regulatory module. *PLoS Biol* **3**, e93 (2005).
178. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
179. Newman, J. R. S. *et al.* Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840–846 (2006).
180. Munsky, B., Neuert, G. & van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science* **336**, 183–187 (2012).
181. Yucel, G. & Small, S. Morphogens: precise outputs from a variable gradient. *Curr Biol* **16**, R29–31 (2006).
182. Bothma, J. P., Levine, M. & Boettiger, A. Morphogen gradients: limits to signaling or limits to measurement? *Curr Biol* **20**, R232–4 (2010).
183. Tabata, T. & Takei, Y. Morphogens, their identification and regulation. *Development* **131**, 703–712 (2004).
184. Chen, H., Xu, Z., Mei, C., Yu, D. & Small, S. A system of repressor gradients spatially organizes the boundaries of Bicoid-dependent target genes. *Cell* **149**, 618–629 (2012).
185. Papatsenko, D. & Levine, M. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the Drosophila embryo. *Proc Natl Acad Sci USA* **102**, 4966–4971 (2005).
186. Manu *et al.* Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. *PLoS Biol* **7**, e1000049 (2009).
187. Jiang, J., Hoey, T. & Levine, M. Autoregulation of a segmentation gene in Drosophila: combinatorial interaction of the even-skipped homeo box protein with a distal enhancer element. *Genes Dev* **5**, 265–277 (1991).
188. Reinitz, J. & Sharp, D. H. Mechanism of eve stripe formation. *Mech Dev* (1995).
189. Harding, K., Hoey, T., Warrior, R. & Levine, M. Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. *EMBO J* (1989).
190. Frohnhöfer, H. G., Lehmann, R. & Nüsslein-Volhard, C. Manipulating the anteroposterior pattern of the Drosophila embryo. *J Embryol Exp Morphol* **97 Suppl**, 169–179 (1986).
191. Driever, W. & Nüsslein-Volhard, C. The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell* **54**, 95–104 (1988).
192. Bradley, R. K., Merkin, J., Lambert, N. J. & Burge, C. B. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol* **10**, e1001229–e1001229 (2012).
193. Zuo, P. *et al.* Activation and repression of transcription by the gap proteins hunchback and Krüppel in cultured Drosophila cells. *Genes Dev* **5**, 254–264 (1991).
194. Sauer, F. & Jäckle, H. Dimerization and the control of transcription by Krüppel. *Nature* **364**, 454–457 (1993).
195. Hoch, M., Schröder, C., Seifert, E. & Jäckle, H. cis-acting control elements for

Kruppel expression in the Drosophila embryo. *EMBO J* **9**, 2587–2595 (1990).
196. Gaul, U., Seifert, E., Schuh, R. & Jäckle, H. Analysis of Krüppel protein distribution during early Drosophila development reveals posttranscriptional regulation. *Cell* **50**, 639–647 (1987).
197. Lott, S. E. *et al.* Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-seq. *PLoS Biol* **9**, e1000590 (2011).
198. Venken, K. J. T. *et al.* Versatile P[acman] BAC libraries for transgenesis studies in Drosophila melanogaster. *Nat Meth* **6**, 431–434 (2009).
199. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* (2009).
200. Zhang, Y., Liu, T., Meyer, C. A. & Eeckhoute, J. Model-based analysis of ChIP-Seq (MACS). *Genome …* (2008).
201. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).
202. Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144–2146 (2011).
203. Gallo, S. M. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *… acids research* (2011).
204. Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **74**, 829–836 (1979).
205. Perry, M. W., Boettiger, A. N. & Levine, M. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proc Natl Acad Sci USA* (2011).doi:10.1073/pnas.1109873108
206. Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol* **12**, R43 (2011).
207. Harrison, M. M., Li, X.-Y., Kaplan, T., Botchan, M. R. & Eisen, M. B. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genet* **7**, e1002266 (2011).
208. Jambunathan, N. *et al.* Multiple bromodomain genes are involved in restricting the spread of heterochromatic silencing at the Saccharomyces cerevisiae HMR-tRNA boundary. *Genetics* **171**, 913–922 (2005).
209. Gradolatto, A. *et al.* A noncanonical bromodomain in the AAA ATPase protein Yta7 directs chromosomal positioning and barrier chromatin activity. *Mol Cell Biol* **29**, 4604–4611 (2009).
210. Lombardi, L. M., Ellahi, A. & Rine, J. Direct regulation of nucleosome density by the conserved AAA-ATPase Yta7. *Proc Natl Acad Sci USA* **108**, E1302–11 (2011).
211. Kurat, C. F. *et al.* Restriction of histone gene transcription to S phase by phosphorylation of a chromatin boundary protein. *Genes Dev* **25**, 2489–2501 (2011).

212. Lin, C. *et al.* Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC). *Genes Dev* **25**, 1486–1498 (2011).
213. Zhang, Y., Shin, H., Song, J. S., Lei, Y. & Liu, X. S. Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq. *BMC Genomics* **9**, 537 (2008).
214. Glover-Cutter, K., Kim, S., Espinosa, J. & Bentley, D. L. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat Struct Mol Biol* **15**, 71–78 (2008).
215. Rahl, P. B. *et al.* c-Myc Regulates Transcriptional Pause Release. *Cell* **141**, 432–445 (2010).
216. Wang, P. *et al.* Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Mol Cell Biol* **29**, 6074–6085 (2009).
217. Wittwer, F., van der Straten, A., Keleman, K., Dickson, B. J. & Hafen, E. Lilliputian: an AF4/FMR2-related protein that controls cell identity and cell growth. *Development* **128**, 791–800 (2001).
218. Tang, A. H., Neufeld, T. P., Rubin, G. M. & Müller, H. A. Transcriptional regulation of cytoskeletal functions and segmentation by a novel maternal pair-rule gene, lilliputian. *Development* **128**, 801–813 (2001).