QUANTITATIVE ANALYSIS OF ENHANCER FUNCTION IN THE DORSAL-VENTRAL PATTERNING GENE NETWORK OF THE *DROSOPHILA* EMBRYO

By

Rupinder Sayal

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Biochemistry and Molecular Biology

2012

ABSTRACT

QUANTITATIVE ANALYSIS OF ENHANCER FUNCTION IN THE DORSAL-VENTRAL PATTERNING GENE NETWORK OF THE *DROSOPHILA* EMBRYO

By

Rupinder Sayal

Enhancers are non-coding regions of DNA that coordinate spatio-temporal regulation of gene expression. These regulatory sequences contain binding sites for sequence-specific transcription factors. Enhancers are instrumental in evolution of novel developmental and morphological features, as well as quantitative expression differences in a given population. In order to understand enhancer function and develop generalizable, predictive models for enhancers, it is important to study how enhancers function at a quantitative level. I first developed a suite of reporter gene vectors, which made quantitative measurements of gene expression from enhancer feasible. This “pHonda” suite of vectors is designed for site-specific integration into fly genome to eradicate position effects, as well as it uses a specific 5’-UTR, which allows for a more diffuse distribution of mRNA, making it more amenable to quantitative studies.

Dorsal-ventral patterning is regulated by a master transcription factor, Dorsal, which is the fly homolog of mammalian NF-κB protein. Dorsal regulates about 100 genes in early fly embryo and coordinates dorsal-ventral patterning. A number of enhancers for these genes have been identified, which have been found to contain binding sites for the above proteins. The availability of several tested enhancer sequences, and quantitative data for concentrations of these

factors, make it a suitable system for carrying out quantitative studies. Using systematic mutagenesis and confocal microscopy I first generated a systematic perturbation dataset for enhancer of *rhomboid* gene. Next, I applied thermodynamic modeling to this dataset, which uses assumptions of statistical thermodynamics to derive gene expression as a function of probabilities of binding of different factors to enhancer sequences, and tested several models for protein cooperativity and repression on this dataset. Subsequently, I used these models to predict gene expression from enhancer sequences, which were not used for modeling, and found that the top-ranked models can predict gene expression from these sequences in a tissue-specific manner. My study highlights the importance of mathematical modeling to understand the general rules of enhancer function.

•

This thesis is dedicated to the two people to whom I owe everything
I have, have had and will have in life,
my father, Mr. Amarjit Sayal and my mother, Mrs. Usha Sayal.

Acknowledgments

I have been extremely fortunate to receive the support and encouragement from my colleagues, mentors and my family during my thesis. First and foremost, I am extremely grateful to my advisor, Dr. David Arnosti, who guided and motivated me in all the different aspects of my graduate training with indefatigable spirit. I want to express my gratitude to my thesis committee members, Dr. C. Titus Brown, Dr. David Dewitt,, Dr. Ian Dworkin, Dr. Laurie Kaguni and Dr. John LaPres for their guidance and advice.

I would like to acknowledge the support of all the previous and past members of Arnosti lab. for helpful discussions, advice and help throughout my thesis. Their presence not only helped in my scientific training, but also made it a great place to work. In particular, I would like to thank Dr. Ahmet Ay, Dr. Walid Fakhouri, Dr. Jacqueline Dresch, Irina Pushel, Benjamin Taylor, Ramona Beckman, Max Winkler, Rewatee Gokhale and Anne Sonnenschein for their assistance in my thesis research.

I am very fortunate to be blessed with a loving family, and without their love and support, this thesis would have never seen the daylight. I would like to thank my parents, my sister Shalini Kaushal and my brother-in-law Puneet Kaushal for making it all worthwhile.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

CMA-ES	Covariance Matrix Adaptation-Evolutionary Strategy
CRM	*cis*-Regulatory Module
DPE	Downstream Promoter Element
DV	Dorso-ventral
eve	even-skipped
MTE	Motif Ten Element
ORF	Open Reading Frame
PB	Pause Button
rho	rhomboid
RMSE	Root Mean Square Error
sna	snail
SV40	Simian Virus 40

CHAPTER I

Introduction

PREFACE

Genes are the fundamental units of heredity. Austrian monk Gregor Mendel first proposed the idea that genes carry information for traits in an organism in the latter part of 19th century. Since then, studies in genetics, biochemistry and molecular biology have provided us with a wealth of information on the process of expression of organismal traits from the genetic information encoded by the genome. Although proteins are the workhorses of a cell, the information for the structure and function of a protein is ultimately derived from the nucleotide sequence of the coding regions of a gene, also known as the open reading frame. In eukaryotes, the multi-subunit enzyme RNA polymerase II transcribes this information into an RNA molecule known as messenger RNA (mRNA). The sequence of nucleotides on the mRNA acts as a template for protein synthesis by the ribosome. The transfer of cellular information from DNA to mRNA to protein constitutes the central dogma of biology, and regulation at each step of this process provides the diversity of gene expression programs required for development and differentiation of organisms, as well as physiological responses to the environment. Transcriptional control constitutes the primary level of control, underlying most biological processes involving gene regulation.

The idea that gene expression is regulated first originated from studies in bacteria, which produce enzymes required for utilization of different sugars based on their availability. Research by Jacob and Monod, as well as other pioneers in this field led to the emergence of ideas of activation as well as repression of gene expression (*1*). In the studies of *lac* operon, it was found that proteins now known generally as transcription factors control gene expression by binding to specific regulatory

sequences of DNA (binding sites). Subsequently, regulatory regions for controlling gene expression were discovered in eukaryotes as well. Viral transcriptional regulatory elements that were first identified up-regulate gene expression by about a thousand-fold, therefore these regions were named “enhancers” (*2*). Later, studies in various model organisms led to the realization that enhancers do not simply activate gene expression; these elements act as integration centers for cues from signaling and developmental pathways, transmitting both positive and negative stimuli to control transcription. Enhancers usually range in size from 200-1000 bp, and have binding sites for multiple transcription factors, which include activators as well as repressors. Multiple activators are commonly found to regulate a given enhancer, employing the principle of activator synergy to stimulate transcription by cooperative interactions (*3*).

Enhancers are involved in most biological processes where differential gene expression is pivotal to organismal function. A slew of studies have characterized enhancers involved in embryonic development, response to environmental stimuli and evolution of novel morphological phenotypes (*4*-*6*). Misregulation of enhancer function by deleterious mutations is associated with various congenital and somatically acquired diseases, which may be induced by various environmental or dietary mutagens (*7*-*9*).

In prokaryotes, *cis*-regulatory regions are generally located 5’ of the transcriptional initiation site of target genes, and transcription factors interact directly with the RNA polymerase. However, in higher eukaryotes, *cis*-regulatory regions have much more complex organization, and can be located at multiple different locations in the genome. Multiple discrete enhancers are found to regulate gene expression either in

same tissue, different tissues, or different temporal stages. These characteristics make the task of identification and characterization of enhancers tedious in metazoans.

1.1 Mechanism of Enhancer Function

Enhancers typically contain binding sites for multiple transcription factors with different biochemical activities. These factors can be either activators, viz., proteins that can positively regulate gene expression, or conversely, repressors. In prokaryotic systems, activators and repressors bind to their recognition sequences and interact directly with components of basal transcription machinery. However, in eukaryotes, transcription factors binding to enhancers act via a variety of mechanisms; these proteins can interact with the basal transcriptional machinery, modify the chromatin template, change localization of the gene within the nucleus, and interact with other transcription factors in complex manners. Modification of genomic DNA and chromatin can be accomplished by recruitment of transcriptional cofactors, which can mediate positive or negative effects.

Activator proteins stimulate gene expression through a variety of mechanisms:

1. Interactions with the basal transcription machinery, including,
    i) Mediator (*10*, *11*)
    ii) TAFs (TBP-associated factors) (*12*, *13*)
    iii) P-TEFb kinase (*14*)
2. Interactions with chromatin modifying complexes (*15*-*17*)
3. Interactions with chromatin remodeling complexes (*18*, *19*)
4. Chromosome dynamics, such as looping via PTS, or subnuclear targeting (*20*, *21*)

A characteristic feature of eukaryotic enhancers is the synergy exhibited by binding of multiple activators. This synergistic effect may be due to several reasons: a) cooperativity at the level of DNA-binding, which is reflected in linked binding sites (*22*, *23*), b) indirect cooperativity, where binding of one activator facilitates nucleosome displacement, allowing other activator to bind to nearby sites in the enhancer (*24*), c) cooperative recruitment of co-activator proteins (*19*, *25*), or d) interaction of activators with different components of basal transcription machinery (*26*). The cooperative interactions observed between eukaryotic activator proteins reflect an essential aspect of the chromatinized transcriptional template in these organisms; ectopic transcription is suppressed by the generally repressive effects of histone proteins interacting with DNA. Only in cases where multiple proteins can work together to drive gene expression from a bona fide regulatory element will transcription occur; single binding events from spurious sites are less likely to stimulate gene expression. Thus the high threshold imposed by chromatin on the activity of cis regulatory elements enables the specificity of gene expression in larger eukaryotic genomes, even as it necessitates more complex biochemical interactions among transcriptional control proteins. The requirement for "team action" amongst these players underlies the essence of the regulatory grammar that this thesis is devoted to understanding.

Eukaryotic repressor proteins are found to generally work through the chromatin-modifying activities of recruited corepressors, which in turn can interfere with the activity of transcriptional activators or the basal machinery. Repressors active in early *Drosophila* embryonic development have been classified into two categories based on the distances at which these can repress nearby activators – short-range e.g., Giant,

Krüppel, Knirps, Snail (up to 100 bp) or long-range e.g. Dorsal and Hairy (up to 1000 bp) (*27*, *28*). Despite having different ranges of action, these repressors recruit common co-repressors, CtBP and Groucho. Chromatin studies have found that the Knirps short-range repressor causes only local, restricted changes in histone density and acetylation levels, while exerting no effects on RNA polymerase density on the target *even-skipped* gene. On the other hand, long-range repressor Hairy does not cause any changes in histone density, but instead causes locus-wide histone deacetylation and abolishes RNA polymerase interaction with the target gene. These studies indicate the diversity of mechanisms used by different repressors in development, while recruiting similar corepressors (*29*).

In addition to different effects on chromatin, different corepressors can also provide diverse mechanisms of repressor action. Mi2-NURD corepressors complex has dual activities of nucleosome remodeling and histone deacetylation (*30*). CtBP corepressor can interfere directly with CBP/p300 activators (*31*). Histone deacetylation is one of the main biochemical activities mediated by a well-known corepressor complex, Sin3, by its interactions with histone deacetylases HDAC1 and HDAC2. This activity is utilized via direct recruitment by repressors like p53 and Elk1, and indirectly by nuclear hormone receptors, where Sin3 is recruited by NCoR and SMRT protein complexes, which then in turn recruit Sin3 (*32*). The diverse array of corepressors available in different cellular contexts is used by the same repressor in a temporal or context-specific fashion. For example, Brinker, a repressor active in developing *Drosophila* imaginal disc, uses Groucho or CtBP, and sometimes neither, to repress different Dpp target genes. In order to understand enhancer function in greater detail, we need to

have a more comprehensive understanding of the variety of cofactors recruited by different DNA-binding proteins and their effect on different classes of genes regulated by the same set of proteins in different contexts.

Transcription factors frequently activate their target genes in response to cues from signaling pathways. One surprising finding from studies of diverse signaling pathways has been that activation and repression of target genes of a pathway are mediated by the same protein using the same DNA sequences or response elements. For example, in Wnt signaling pathway, beta-catenin, which is the central activator protein in this pathway, is phosphorylated and degraded by ubiquitin pathway. Tcf/Lef protein represses target genes in absence of Wnt signaling by associating with Groucho, which recruits histone deacetylases to mediate repression (*33*). However, in response to Wnt signal, the kinase is inhibited and beta-catenin is not degraded. Beta-catenin accumulates, translocates into nucleus, and forms a complex with transcription factor Tcf/Lef to activate target genes (*34*). This example is one of many documented in several different signaling pathways, which illustrate how key regulators of these pathways switch the repression and activation activities to modulate gene expression (*35*). Such dual usage of binding sites to mediate both transcriptional repression and activation indicates that without knowledge of signaling states of a cell, the sequence information embedded in the DNA is not a conclusive guide to the transcriptional readout of the genome. The concept is not new, having been explored as far back as the first studies of the *lac* operon, but the complexity is one aspect of enhancer "grammar" that this thesis does not address. I discuss the implications of the integration of signaling information and DNA-based transcriptional control information in Chapter 4.

1.2 Enhancers and Evolution

Variation in enhancers appears to be a primary source of phenotypic diversity between related species and may account for a large amount of the expression differences between species (*24*, *36*). Expression differences can be due to *cis*- or *trans*- effects. In a study where researchers used RNA-seq to quantify *cis*- and *trans*-divergence expression differences between two related fly species, *Drosophila melanogaster* and *Drosophila sechellia* in parental species and hybrids, it was found that 78% of expressed genes showed evidence of expression divergence. The relative contributions of *cis*- and *trans*-regulatory divergence to expression differences were 51% and 66% respectively (*19*, *25*, *37*). A similar study was done in two related yeast species, *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*, which had diverged from their last common ancestor about 5 million years ago. Microarrays were used to compare allele-specific expression in parental species as well as hybrids, under four different growth conditions. The researchers found that in three of the four conditions (heat shock, rich media and addition of a histone deacetylase inhibitor, Trichostatin A), *cis*-effects were dominant (*26*, *38*). These studies suggest that *cis*-regulatory changes can account for a large amount of expression differences between species and may be responsible for speciation, phenotypic evolution as well as development of novel morphological features. In addition, trans effects may reflect differential expression of transcription factors, which involve changes in the regulatory regions controlling these proteins (*27*, *28*, *39*).

Recently, numerous studies have revealed how changes in enhancers and their target gene expression levels or spatio-temporal patterns have led to evolution of

distinct morphological features during evolution. In a study of the larger limb of bats as compared to other mammals, researchers found that the differences could be traced to a specific enhancer of a transcription factor, Paired-related homeobox gene 1 (*Prx1*), an important regulator of skeletal limb elongation. Replacement of mouse *Prx1* enhancer by an equivalent bat sequence resulted in higher transcript levels for *Prx1* in transgenic mice. Consequently, these transgenic mice exhibited larger forelimbs, recapitulating the phenotype of the bats (*29*, *40*).

Convergent evolutionary changes are also found to involve adaptation of expression patterns of common targets. Mimicry by wing colors and patterns has evolved multiple times in a genus of butterflies widely distributed in South America, *Heliconius*. Expression-analysis and genome-wide association studies have revealed the causal factor to be an intron-less homeobox transcription factor, *optix,* whose expression was found to 'prefigure' wing patterns. Although this study didn't find any specific *cis*-regulatory elements responsible for wing patterns, it is likely that a variety of wing patterns can be generated by loss and gain of specific enhancer elements (*30*, *41*).

An interesting example of how fine-tuned changes in enhancer structure can lead to divergent phenotypes in a sex-specific manner comes from studies on abdominal pigmentation in flies. In *Drosophila*, males have pigmented dorsal cuticular plates (tergites), in the abdominal segments A5 and A6, which are the posterior-most; whereas in females, this pigmentation extends only to A2-A4, which is common to both sexes. Genetic studies have indicated that a transcription factor, *bric-a-brac* (*bab*), controls this phenotype, and is itself regulated by two other transcription factors, Abdominal-B (Abd-B) and Doublesex (DSX). BAB is a repressor, and researchers showed that in females,

Abd-B and female-specific isoform of DSX, called DSX^F activated BAB expression, leading to loss of pigmentation in A5-A6 segments, whereas in males, male-specific isoform of DSX, known as DSX^M, repressed BAB expression, leading to pigmentation (*31*, *36*). Similar results have been found from analysis of lineage-specific trichome patterns in two fly species (*32*, *42*), and disappearance of pelvic spines in stickleback fish(*43*). The rapidly growing field of evolutionary developmental biology ('evo-devo') has cataloged a large number of such studies where regulatory changes in enhancers have repeatedly led to morphological changes during evolution (*44*). But, how are changes in enhancers tolerated while subtly leading to new phenotypes in the course of thousands of generations? This question inevitably leads to how much "leg-room" is there for an enhancer, which allows it to incorporate sequence changes while still maintaining function. This question is inevitably tied to the question of spatial relationships between binding sites for various proteins on an enhancer, which is discussed in detail in next section.

1.3 Enhancer grammar

Enhancers generally contain multiple binding sites of varying affinities for diverse transcription factors. For a particular enhancer, the arrangement, number, and quality of binding sites are distinct from other enhancers that may be bound by the same regulatory proteins. These differences may be trivial, or they may have measureable impact on the functional output of the element. The contribution of the internal arrangement and properties of binding sites on enhancer to functional output has been referred to as the "enhancer grammar" (*45*).

The Drosophila blastoderm embryo has provided numerous examples of the types of sequence-dependent transcriptional readout that points to such enhancer grammar. Dorsal, Twist and Snail are three transcription factors that control the expression of about 100 genes in the early fly embryo (*46*). Although the same proteins are involved in regulation of the target genes, the patterns of these genes' expression vary markedly. For instance, in the lateral neuroectodermal region, the widths of expression range from 10 nuclei for *brk* (*brinker*) to 6 nuclei for *vnd* (*ventral nervous system-defective*) (*39*). The differences suggest that particular features of the enhancers involved influence interactions between transcription factors, and thus functional distinctions. The spatial relationship between Dorsal and Twist binding sites is one source of this variation. In a study involving genes expressed in neurectoderm of early fly embryo, it was observed that the broadest stripes of expression were correlated with 7-12 bp spacer between Dorsal site and nearest E-box site, while those deviating away from this distance had narrower stripes of expression. Changing the sub-optimal distance of binding sites in *vn* to the optimal one by insertions led to expansion of expression pattern, very similar to broad pattern of *sog* (*39*). This study highlighted the importance of how minor changes in distances between binding sites can lead to widely different patterns of gene expression in the same lineage.

This study also examined expression of these neurectodermal genes in two other fly species, *D. pseudoobscura* and *D. virilis*. The expression patterns driven by putative enhancers of *rho*, *vn*, *vnd*, and *brk* were wider for *D. virilis* and narrower for *D. pseudoobscura* when these enhancers were tested in *D. melanogaster*. However, the endogenous expression patterns for these genes looked identical in these three

lineages. This result suggests that concomitant to enhancer sequence variation, changes in the levels or activity of the trans-acting factors e.g., Dorsal and Twist influence the readout of these enhancers. The authors suggest that changes in protein levels may drive compensatory changes in the sequences of enhancers, highlighting the importance of the trans-regulatory field when discerning the enhancer grammar that is read out by these proteins.

Here, I discuss three models that present different views of enhancer grammar; the range of ideas may not be exhaustive, and the models may not be mutually exclusive.

1) The Enhanceosome Model - The Interferon-beta enhancer is a prototypical example of the 'enhanceosome', and it was the research on this system that has largely driven the concept of the enhanceosome. This short 50 bp enhancer binds three different transcriptional activator protein complexes through motifs specific for NF-κB, IRF-3/IRF-7 and ATF-2/c-Jun (*47*). None of these DNA elements or proteins can activate transcription on their own. Multiple copies of individual DNA elements can activate transcription, but at much lower levels than an intact enhancer. The intact enhancer responds specifically only to virus infection, but individual elements can respond to other inducers as well. Thus, various parts of enhancer, when combined as a whole, control the expression level as well as specificity of induction of target gene. This enhancer has a strict requirement for tight linkage of binding sites, which are needed for cooperative assembly of a multi-protein complex ('enhanceosome') (*47*).

2) The Billboard Model - In contrast to the enhanceosome picture of enhancer function in which there is a low tolerance for variation in positioning and number of binding sites,

many enhancers acting on developmental genes of the early fly embryo work more like "flexible billboards" with a high degree of flexibility of positioning and number of binding sites. Loss of particular activator sites can be compensated by new sites in ectopic locations (*48*). In particular cases, different subsets of binding sites on a small element can display contrasting information to the cellular transcriptional machinery. These billboard-type enhancers are not infinitely flexible, for further changes in arrangements and stoichiometries of binding sites can result in different transcriptional outputs (*49*).

3) The Transcription Factor Collective Model - A third model for enhancer grammar has emerged from study of CRMs involved in gene expression in cardiac mesoderm (CM) and visceral mesoderm (VM) in early *Drosophila* development. Hundreds of enhancers involved in this process are bound by five key transcription factors: pMad, dTCF, Doc, Pnr, and Tin. For a subset of these elements, the researchers did not find a correlation between in vivo protein occupancy and predicted binding sites for many of these transcription factors. The researchers suggested that many or most of the promoter-enhancer targeting takes place at the level of protein-protein interactions (*50*). This study raises important questions about the mechanism of assembly of proteins on an enhancer; if most of the specificity is not directly through protein-DNA contacts, attempts to understand function of these elements through simple analysis of DNA sequences will be a fraught experience. Although it remains to be seen how widely applicable this model is, it implies that underlying sequence may have little direct predictive power for inferring regulatory information, and consequently, enhancer validation will need to be carried out individually, a tedious and time-consuming process

involving systematic testing and characterization of enhancer sequences using reporter gene assays.

These models have emerged after analyses of a few enhancers involved in different biological contexts. These enhancers bind different proteins, which themselves recruit cofactors harboring divergent biochemical activities. It is likely that these additional layers of complexity also drive the evolution of enhancer structure, which is readily seen in the case of IFN-β enhancer, which has hardly undergone any sequence turnover due to requirements of rigid linking of binding sites. The above models of enhancer function may not be mutually exclusive, and endogenous enhancers may have features from different models. We need a systematic survey of contributions of binding sites and incorporate natural variability from a wide variety of enhancers active in different contexts to understand enhancer function on a molecular level.

1.4 Dorsal-ventral patterning gene enhancers

The elucidation of enhancer function has come largely from reverse engineering of endogenous regulatory elements. One of the richest sources of knowledge about enhancer structure and function has been the Drosophila blastoderm embryo, which is patterned during early stages of development by the action of diverse transcriptional cascades that are responsible for generating primary positioning information. One key method by which this is achieved is through the action of morphogens. Morphogens are substances distributed in a graded fashion in a developmental field, and exposure of cells to different concentrations of morphogens causes cells to assume diverse developmental fates. Two morphogens set up the body plan for a developing fly embryo in the two principal axes. The Bicoid transcription factor regulates genes involved in

anterior-posterior patterning, whereas the Dorsal transcription factor is the morphogen responsible for dorsal-ventral patterning. Dorsal belongs to the Rel family of transcription factors, the members of which include mammalian NF-κB and c-Rel (*51*, *52*). Dorsal is activated by the Toll pathway, which involves signaling through an IL-1 receptor to release cytoplasmic Dorsal protein to the nucleus, and different concentrations of Dorsal protein in different positions of the embryo drive differentiation of three primary germ layers, viz., mesoderm, neurogenic ectoderm and non-neurogenic (dorsal) ectoderm (*53*).

As mentioned above, Dorsal activates and represses a total of about 100 genes in the developing *Drosophila* embryo (*46*). Two of the earliest genes activated by Dorsal are *twist* (*54-56*) and *snail* (*57*). Snail and Twist proteins are the primary determinants of mesoderm differentiation in fly embryo (*58*). The zinc-finger protein Snail is a transcriptional repressor, and acts over short-ranges of up to 100 bp to interfere with ("quench") activators within a distance of about 100 bp (*59*). Twist is a transcriptional activator, belonging to the helix-loop-helix family of proteins (*60*). *In vitro* gel retardation assays have shown that Dorsal and various bHLH factors can interact, suggesting that these proteins may cooperate synergistically to activate gene expression in neurectoderm, where there are limiting amounts of Dorsal and Twist (*61*).

Researchers have identified more members of this regulon by comparing changes in gene expression in different mutant backgrounds of Toll-receptor signaling, leading to identification of about 40 new Dorsal target genes (*62*). Improvements in technology made available by genome-wide tiling arrays helped to identify 26 more Dorsal target genes (*63*).

A further advance in our understanding of how dorsal-ventral patterning network is wired came from genome-wide binding maps for Dorsal, Twist and Snail obtained using ChIP-chip (Chromatin Immunoprecipitation followed by microarray analysis). This study identified many new putative Dorsal-regulated enhancers, including many for genes involved in Dpp signaling as well as anterior-posterior patterning, indicating integration of diverse signaling pathways and patterning processes in the early fly development (*46*).

One of the surprising finding of previous study was the observation that many Dorsal target genes appear to have more than one enhancer for identical expression patterns. Subsequent studies revealed that the phenomenon of more than one enhancer for Dorsal target genes was more widespread than previously thought. It was postulated that these secondary enhancers, sometimes termed “shadow” enhancers, were evolving faster than principal enhancer and may be contributing to animal diversity by evolving a different function while the principal enhancer can still maintain the original expression pattern of the gene (*64*).

Along with Dorsal, Twist and Snail, another protein, Zelda, was found to be important for activation of zygotic genes in the early *Drosophila* embryo during maternal-to-zygotic transition (*65*, *66*). Visualization of gene expression of many zygotically active, developmental patterning genes in $zld^{-/-}$ embryos showed these genes had aberrant or weak expression, pointing to importance of Zelda in activating these genes, including many involved in dorsal-ventral patterning. Zelda binds to motifs known as TAGteam motifs containing the trinucleotide TAG (*67*). *In vitro* experiments showed that Zelda and Grainyhead, another repressor protein, compete with each other

to bind to enhancer regions of *dpp* and *tll*. This competition was postulated to be important for temporal coordination of expression of these genes (*68*). Subsequent ChIP-seq studies of Zelda binding in early fly embryo showed that the protein was important for regulating expression of many more genes than previously thought, including precise and robust activation of many patterning genes. Also, Zelda binding overlaps with previously characterized genomic binding "hot spots" for binding of numerous developmental transcription factors (*69*). This observation led the authors to postulate that Zelda coordinates the formation of these hotspots, a speculation also made by some other researchers gazing at these genome-wide binding maps (*70*, *71*).

One of the genes regulated by the protein trio of Dorsal, Twist and Snail is *rhomboid (rho). rho* was identified by genetic screens as one of the genes involved in dorsal-ventral patterning (*72*). The *rho* gene encodes an intramembrane serine protease involved in EGFR signaling (*73*). *rho* mutant embryos display both peripheral nervous system (PNS) and musculature defects. In larvae, two out of five lateral chordotonal organs (stretch receptors) are missing. In adult stage, *rho* mutant flies also show an abnormal muscle pattern. Thoracic segments have altered and irregular muscle patterns, and some muscle fibers are affected in abdominal segments as well (*74*). Traditional enhancer testing using reporter genes identified a regulatory segment responsible for neuroectoderm expression located upstream of the *rho* basal promoter. DNase I footprinting experiments have revealed that Dorsal, Twist and Snail have four, two and four binding sites respectively, in this 300-bp enhancer, referred to as the *rho* neurectoderm enhancer (*rho*NEE, Fig. 1.1). Experiments in which *lacZ* expression was driven using various versions of the *twi* promoter with different distances separating

linked Dorsal and E-box (putative *twi*-binding) sites showed that closely linked sites are essential for neurectoderm expression, and this linkage is dispensable for mesoderm expression, where there are high levels of Dorsal and Twist (*75*). Snail is the repressor of *rho* expression in mesoderm, and removal of Snail-binding sites on *rho* enhancer leads to de-repression (expansion) of expression into mesoderm.

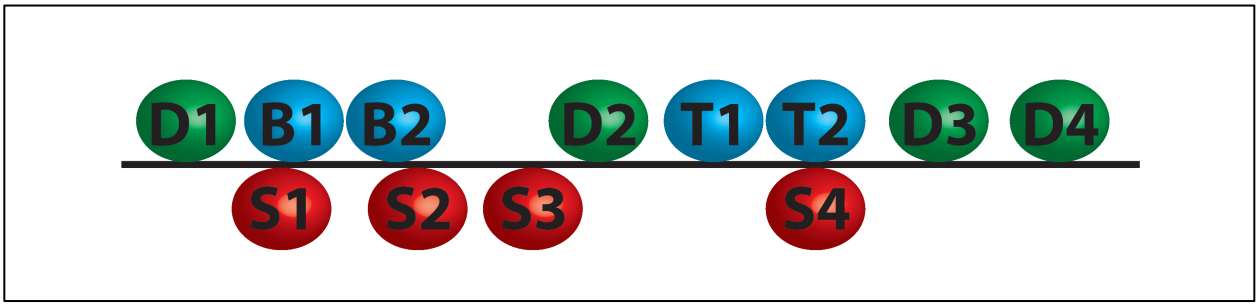


Figure 1.1 A schematic illustration of *rhomboid* Neurectodermal Enhancer (*rho*NEE). The *rho*NEE has binding sites for Dorsal, Twist, Snail and bHLH factors, indicated by letters D, T, S and B respectively. The numbers with binding sites indicate their order from 5' to 3' on the 318 bp enhancer.
(For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation).

1.4 Thermodynamic mathematical modeling of enhancer function

In the preceding sections, I stressed the importance of characterization of enhancer sequences to understand gene regulation in metazoans. Studies in the past 30 years since the identification of first enhancer sequences have given us a rich, detailed list of enhancer sequences in different organisms, involved in myriad biological processes. We know that various DNA-binding- as well as non-DNA-binding proteins play important roles in gene expression by interacting with specific recognition sequences present in enhancer sequences. The combinatorial output of these

interactions determine the ‘state’ of gene expression in different cells of an organism, where these outputs may be different due to presence or absence of different proteins, giving rise to different cell fates.

In spite of this vast body of knowledge regarding enhancers, we are far from a mechanistic and definitive picture of enhancer function. Bioinformatic predictions are of limited utility, because the binding sites of most transcription factors are present in genome in large numbers due to their degeneracy, and many of these sites are not actually bound in vivo. Direct biochemical measurements of in vivo occupancy too have their limits; genome-wide studies of binding of transcription factors have revealed widespread binding of these proteins to thousands of genomic regions, of which it is believed that many are likely to be non-functional. Another challenge is presented by the evidence of contribution of enhancer sequences to phenotypic changes, leading to morphological evolution. How do these changes modulate enhancer function? Since the changes occur over millennia, and are a result of step-wise changes to enhancer sequences, how do these changes get selected and retained over many hundreds of generations? These questions can be approached experimentally to some degree, but reporter assays suffer from several limitations: first, the sequences are tested out of their genomic context; second, the assays are costly and laborious, and third, most of these assays are qualitative in nature.

The preceding questions reveal the limits of our knowledge regarding enhancer function, and the need to transition from a qualitative to quantitative understanding of enhancer structure and function. Instead of a cursory picture of enhancers, where a minimal enhancer is isolated, cloned and tested using reporter gene assays in cultured

cells, and mutations to key binding sites abolish expression, we need a new paradigm where we can measure the contribution of putative enhancer sequences in terms of individual binding sites, and calculate how changes in those sites may lead to aberrant expression, and addition of sites for new regulators may lead to rewiring of regulatory networks, which in turn may cause appearance of new morphological phenotypes. Thermodynamic models represent one of the approaches where transcriptional outputs from enhancer sequences can be gauged in a quantitative manner.

Thermodynamic modeling of enhancer function is based on the ideas of statistical mechanics that the gene expression from any given regulatory DNA sequence is a function of the individual contributions of each of the "states" that the enhancer can assume. The overall expression from an enhancer can be calculated by estimating the equilibrium probabilities of binding of different transcription factors based on their concentration and binding affinities (*76*, *77*). Such models incorporate the effects of binding of multiple proteins, which can be activators or repressors, and cooperative or antagonistic interactions. The statistical framework can be theoretically applied to any regulatory DNA of interest, containing any number of binding sites for various proteins. In the past, these models have been successfully applied to modeling gene regulation in bacteria (*78*), as well as in yeast (*79*) and *Drosophila* (*80-83*).

Our laboratory took a systematic approach in a recent study to decipher the parameters of repression and cooperativity in fly enhancers. We constructed synthetic enhancer constructs with binding sites for activators Dorsal and Twist, along with sites for short-range repressors Giant, Kruppel and Knirps, in various configurations where distance, number and arrangement of repressor binding sites was changed, while

keeping the position and number of activator sites constant. We used confocal microscopy to measure repressor protein concentration and *lac*Z mRNA driven by enhancer constructs and global, 'evolutionary strategy' technique of parameter estimation to calculate parameters for scaling factors for repressors, repressor cooperativity and distance-dependent quenching of activators by repressors. Our study found modest levels of repressor cooperativity, which differed from earlier estimates, along with similar quenching parameters for activators, indicating that short-range repressors exert similar biochemical effects on different activators. We found a surprising non-monotonic, distance-dependent quenching function for repressors (*82*). This might indicate that Giant may have a preferred distance to nearby activators for maximum repression efficiency.

The different mechanistic predictions put forth in this study lead to specific hypotheses that can be tested using experimental approaches, as well as validated using comparisons of the entire set of enhancers regulated by the same set of proteins. This study used a bottom-up approach for deciphering enhancer structure with a fairly low number of variables in terms of number of proteins, binding site sequence, arrangement and affinity, which allowed for accurate parameter estimation, in contrast with studies involving endogenous enhancer sequences, where the landscape of binding sites is relatively more complex. However, this study highlights the holes in our knowledge in terms of complex nature of enhancer sequences and the necessity of datasets with quantitative outputs for all the factors that may affect enhancer function.

Thermodynamic models were applied in another study on the structure of enhancers of genes regulated by same set of proteins. Here, researchers used confocal

imaging to quantify the protein levels of Dorsal, Twist and Snail in early fly embryo. The researchers focused on enhancers of two genes regulated by these proteins, *rhomboid* and *ventral nervous system defective* (*vnd*), which are expressed as stripes of varying thickness in putative ventral neurectoderm. A distinguishing feature in these enhancers is the linkage of Dorsal and overlapping Twist/Snail sites, which the authors referred to as a DTS module. The expression patterns of these genes were measured by using minimal enhancer-*lac*Z fusion reporter genes, as well as endogenous mRNA expression patterns visualized using in situ RNA hybridization. Thermodynamic models were used to fit these expression patterns, and three key features of enhancers were analyzed – number of binding sites, number of DTS modules, and cooperativity between proteins. 128 different enhancer structures, viz., enumeration of binding sites and modules in enhancers, were fit to expression patterns of *rho* and *vnd*, and structures were retained if they fit the observed expression patterns. The parameters for binding constants, cooperativity and number of modules were then analyzed and compared with known enhancers of these genes.

Through analysis of parameters observed for enhancer structures with good fits to expression patterns, as well as observed structures of orthologous enhancers for these two genes from seven fly species, researchers postulated a few features in the enhancers that could be responsible for these differences. Firstly, the *rho* enhancer has only one linked DTS module, whereas the *vnd* enhancer has two. This additional module may impart greater activation potential to the enhancer. Secondly, *rho* enhancer possibly has higher Dorsal-Twist cooperativity as well as higher Twist-Twist cooperativity. Thirdly, *vnd* enhancer has higher Snail-Snail cooperativity (*80*). A caveat

to these assertions is that these are based on bioinformatics-based predictions of binding sites, and lack empirical evidence based on experimental manipulation and mutation of binding sites to show that these sites are indeed the causal factors for difference in expression patterns.

Further, a closer look at the model using sensitivity analysis raised some doubts about the conclusions. Sensitivity analysis can be defined as a measure of effect of change in values of individual parameters on model output. More sensitive parameters lead to greater changes in model output. So, sensitive parameters impart us with more confidence in relating parameter values to biological insights, and are thus more desirable. Global sensitivity analysis on this model revealed that scaling factors are much more sensitive as compared to the cooperativity parameters, which were thought to be the causal factor for divergent outputs of *vnd* and *rho* enhancers. In addition, when a synthetic dataset was tested against experimental dataset and parameters were extracted which gave good correlations, although in 39% of instances, all relationships were found be exactly the same as found in the study, a similar number of parameter combinations were found which gave good correlations, while having drastically different biological relationships among proteins. These results make us reconsider the strength of conclusions drawn by this study.

In a study involving thermodynamic mathematical modeling of early *Drosophila* segmentation, researchers used 44 previously tested enhancers (*84*), along with quantitative profiles of 8 key transcription factors - Bicoid (BCD), Hunchback (HB), Caudal (CAD), Kruppel (KR), Giant (GT), Torso-response element (TorRE), Knirps (KNI), and Tailless (TLL), and asked if the model could predict the expression of these

enhancers. This model used cooperativity between factors that decreased with distance. No thresholds for Position-specific scoring matrices (PSSMs) were used to eliminate weak binding sites. This model had three free parameters: a) factor concentration *in vivo*, b) increased rate of transcription by interaction of the factor and basal transcription machinery, and c) factor cooperativity. In this study, weak sites were found to be important for the predictive power of this model. Also, cooperativity between factors was found to be important for maintaining sharp boundaries of gene expression. The authors of this study also found no evidence for heterotypic clustering of sites as well as overlap of binding preferences of different factors, where the latter may indicate competition between binding sites as a primary mode of enhancer function. The researchers then tested and validated the model parameters on 26 other modules that were not used for model training.

Although this study was instrumental in setting up the stage for using mathematical techniques for dissecting enhancer function, the dataset was qualitative in nature, based on non-quantitative in situ hybridization data. Also, the study was based on enhancers regulated by several different regulators. This thins out the dataset considerably if we consider the expression and sequence data for each regulator. As we have seen for Zinzen *et al.* study, robust parameter estimation requires fairly large datasets with sensitive parameters. Also, Fakhouri *et al.*'s study highlights the intricate nature of landscape of binding sites on enhancers necessitating the construction of large datasets testing various variables exhaustively to tackle the problems of parameter compensation and parameter sensitivity.

These problems collectively emphasize the importance of construction of large, quantitative datasets based on perturbation of a set of enhancers regulated by a tractable number of proteins in a limited number of number and arrangements, and using parameter estimation approaches to derive biologically meaningful insights about enhancer function in general and in a quantitative manner. I describe in the following chapters my approach to address these concerns, and develop a new level of thermodynamic modeling analysis based on extensive and deep quantitative perturbation analysis of well-characterized enhancer sequences regulated by Dorsal, Twist and Snail. I show that this approach, which was carried out primarily in collaboration with my mathematician colleague Jacqueline Dresch, strongly supports the notion that there is a tractable “enhancer grammar” for this system, and that my studies provide a guide to the entire regulon of the Dorsal, Twist, and Snail factors. Furthermore, my studies demonstrate the limitations and opportunities that similar work can anticipate in applying mathematical modeling to transcriptional analysis of genomes.

REFERENCES

REFERENCES

1. B. Muller-Hill, *The Lac Operon: A Short History of a Genetic Paradigm* (Walter De Gruyter Inc, 1996).

2. J. Banerji, S. Rusconi, W. Schaffner, Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences, *Cell* **27**, 299–308 (1981).

3. W. Lee, A. Haslinger, M. Karin, R. Tjian, Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40, *Nature* **325**, 368–372 (1987).

4. M. Levine, Transcriptional Enhancers in Animal Development and Evolution, *Current Biology* **20**, R754–R763 (2010).

5. P. J. Wittkopp, G. Kalay, Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence, *Nat Rev Genet* **13**, 59–69 (2012).

6. C.-T. Ong, V. G. Corces, Enhancer function: new insights into the regulation of tissue-specific gene expression, *Nat Rev Genet* **12**, 283–293 (2011).

7. I. Williamson, R. E. Hill, W. A. Bickmore, Enhancers: From Developmental Genetics to the Genetics of Common Human Disease, *Developmental Cell* **21**, 17–19 (2011).

8. D. J. Kleinjan, P. Coutinho, Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease, *Briefings in Functional Genomics and Proteomics* **8**, 317–332 (2009).

9. N. J. Sakabe, D. Savic, M. A. Nobrega, Transcriptional enhancers in development and disease, *Genome Biol* **13**, 238 (2012).

10. J. L. Stevens *et al.*, Transcription control by E1A and MAP kinase pathway via Sur2 mediator subunit, *Science* **296**, 755–758 (2002).

11. G. T. Cantin, J. L. Stevens, A. J. Berk, Activation domain-mediator interactions promote transcription preinitiation complex assembly on promoter DNA, *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12003–12008 (2003).

12. C. J. Thut, J. L. Chen, R. Klemm, R. Tjian, p53 transcriptional activation mediated by coactivators $TAF_{II}40$ and $TAF_{II}60$, *Science* **267**, 100–104 (1995).

13. F. Sauer, S. K. Hansen, R. Tjian, Multiple TAF_{II}s directing synergistic activation of transcription, *Science* **270**, 1783–1788 (1995).

14. S. R. Eberhardy, P. J. Farnham, Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter, *J. Biol. Chem.* **277**, 40156–40162 (2002).

15. S. R. Bhaumik, T. Raha, D. P. Aiello, M. R. Green, In vivo target of a transcriptional activator revealed by fluorescence resonance energy transfer, *Genes & Development* **18**, 333–343 (2004).

16. M.-H. Kuo, E. vom Baur, K. Struhl, C. D. Allis, Gcn4 Activator Targets Gcn5 Histone Acetyltransferase to specific promoters independently of transcription, *Molecular Cell* **6**, 1309–1320 (2000).

17. E. J. Stockinger, Y. Mao, M. K. Regier, S. J. Triezenberg, M. F. Thomashow, Transcriptional adaptor and histone acetyltransferase proteins in Arabidopsis and their interactions with CBF1, a transcriptional activator involved in cold-regulated gene expression, *Nucleic Acids Research* **29**, 1524–1533 (2001).

18. N. Yudkovsky, C. Logie, S. Hahn, C. L. Peterson, Recruitment of the SWI/SNF chromatin remodeling complex by transcriptional activators, *Genes & Development* **13**, 2369–2374 (1999).

19. G. J. Narlikar, H.-Y. Fan, R. E. Kingston, Cooperation between complexes that regulate chromatin structure and transcription, *Cell* **108**, 475–487 (2002).

20. T. Tsukamoto *et al.*, Visualization of gene activity in living cells, *Nat. Cell Biol.* **2**, 871–878 (2000).

21. T. Tumbar, A. S. Belmont, Interphase movements of a DNA chromosome region modulated by VP16 transcriptional activator, *Nat. Cell Biol.* **3**, 134–139 (2001).

22. A. D. Johnson *et al.*, λ Repressor and cro—components of an efficient molecular switch, *Nature* **294**, 217–223 (1981).

23. S. Oehler, Induction of the lac promoter in the absence of DNA loops and the stoichiometry of induction, *Nucleic Acids Research* **34**, 606–612 (2006).

24. M. A. Schwabish, K. Struhl, The Swi/Snf complex is important for histone eviction during transcriptional activation and RNA polymerase II elongation in vivo, *Molecular and Cellular Biology* **27**, 6987–6995 (2007).

25. G. Wang *et al.*, Mediator Requirement for Both Recruitment and Postrecruitment Steps in Transcription Initiation, *Molecular Cell* **17**, 683–694 (2005).

26. V. M. Weake, J. L. Workman, Inducible gene expression: diverse regulatory mechanisms, *Nat Rev Genet* **11**, 426–437 (2010).

27. S. Payankaulam, L. M. Li, D. N. Arnosti, Transcriptional Repression: Conserved and Evolved Features, *Current Biology* **20**, R764–R771 (2010).

28. A. J. Courey, S. Jia, Transcriptional repression: the long and the short of it, *Genes & Development* **15**, 2786–2796 (2001).

29. L. M. Li, D. N. Arnosti, Long- and short-range transcriptional repressors induce distinct chromatin States on repressed genes, *Current Biology* **21**, 406–412 (2011).

30. I. Zamir, J. Zhang, M. A. Lazar, Stoichiometric and steric principles governing repression by nuclear hormone receptors, *Genes & Development* **11**, 835–846 (1997).

31. J.-H. Kim, E.-J. Cho, S.-T. Kim, H.-D. Youn, CtBP represses p300-mediated transcriptional activation by direct association with its bromodomain, *Nat. Struct. Mol. Biol.* **12**, 423–428 (2005).

32. R. A. Silverstein, K. Ekwall, Sin3: a flexible regulator of global gene expression and genome stability, *Curr. Genet.* **47**, 1–17 (2005).

33. R. A. Cavallo *et al.*, Drosophila Tcf and Groucho interact to repress Wingless signalling activity, *Nature* **395**, 604–608 (1998).

34. C. Y. Logan, R. Nusse, The Wnt signaling pathway in development and disease, *Annu. Rev. Cell Dev. Biol.* **20**, 781–810 (2004).

35. S. Barolo, J. W. Posakony, Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling, *Genes & Development* **16**, 1167–1181 (2002).

36. S. B. Carroll, Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution, *Cell* **134**, 25–36 (2008).

37. C. J. McManus *et al.*, Regulatory divergence in Drosophila revealed by mRNA-seq, *Genome Res.* **20**, 816–825 (2010).

38. I. Tirosh, S. Reikhav, A. A. Levy, N. Barkai, A yeast hybrid provides insight into the evolution of gene expression regulation, *Science* **324**, 659–662 (2009).

39. J. Crocker, Y. Tamori, A. Erives, Evolution acts on enhancer organization to fine-tune gradient threshold readouts, *PLoS Biol* **6**, e263 (2009).

40. C. J. Cretekos *et al.*, Regulatory divergence modifies limb length between mammals, *Genes & Development* **22**, 141–151 (2008).

41. R. D. Reed *et al.*, optix drives the repeated convergent evolution of butterfly wing pattern mimicry, *Science* **333**, 1137–1141 (2011).

42. N. Frankel *et al.*, Morphological evolution caused by many subtle-effect substitutions in regulatory DNA, *Nature* **474**, 598–603 (2011).

43. Y. F. Chan *et al.*, Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer, *Science* **327**, 302–305 (2010).

44. S. B. Carroll, *Endless Forms Most Beautiful: The New Science of Evo Devo* (W. W.

Norton & Company, 2006).

45. M. M. Kulkarni, D. N. Arnosti, cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster, *Molecular and Cellular Biology* **25**, 3411–3420 (2005).

46. J. Zeitlinger *et al.*, Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo, *Genes & Development* **21**, 385–390 (2007).

47. E. Ford, D. Thanos, The transcriptional code of human IFN-beta gene expression, *Biochim. Biophys. Acta* **1799**, 328–336 (2010).

48. D. N. Arnosti, S. Barolo, M. Levine, S. Small, The eve stripe 2 enhancer employs multiple modes of transcriptional synergy, *Development* **122**, 205–214 (1996).

49. M. M. Kulkarni, D. N. Arnosti, Information display by transcriptional enhancers, *Development* **130**, 6569–6575 (2003).

50. G. Junion *et al.*, A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History, *Cell* **148**, 473–486 (2012).

51. R. Steward, Dorsal, an embryonic polarity gene in Drosophila, is homologous to the vertebrate proto-oncogene, c-rel, *Science* (1987).

52. Y. T. Ip, R. Kraut, M. Levine, C. A. Rushlow, The dorsal morphogen is a sequence-specific DNA-binding protein that interacts with a long-range repression element in Drosophila, *Cell* **64**, 439–446 (1991).

53. B. Moussian, S. Roth, Dorsoventral Axis Formation in the Drosophila Embryo—Shaping and Transducing a Morphogen Gradient, *Current Biology* **15**, R887–R899 (2005).

54. D. J. Pan, J. D. Huang, A. J. Courey, Functional analysis of the Drosophila twist promoter reveals a dorsal-binding ventral activator region, *Genes & Development* **5**, 1892–1901 (1991).

55. C. Thisse, F. Perrin-Schmitt, C. Stoetzel, B. Thisse, Sequence-specific transactivation of the Drosophila twist gene by the dorsal gene product, *Cell* **65**, 1191–1201 (1991).

56. J. Jiang, D. Kosman, Y. T. Ip, M. Levine, The dorsal morphogen gradient regulates the mesoderm determinant twist in early Drosophila embryos, *Genes & Development* **5**, 1881–1891 (1991).

57. Y. T. Ip, R. E. Park, D. Kosman, K. Yazdanbakhsh, M. Levine, dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo, *Genes & Development* **6**, 1518–1530 (1992).

58. D. Kosman, Y. T. Ip, M. Levine, K. Arora, Establishment of the mesoderm-neuroectoderm boundary in the Drosophila embryo, *Science* **254**, 118–122 (1991).

59. S. Gray, P. Szymanski, M. Levine, Short-range repression permits multiple enhancers to function autonomously within a complex promoter, *Genes & Development* **8**, 1829–1838 (1994).

60. C. Murre, P. S. McCaw, D. Baltimore, A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins, *Cell* **56**, 777–783 (1989).

61. J. Jiang, M. Levine, Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen, *Cell* **72**, 741–752 (1993).

62. A. Stathopoulos, M. Van Drenth, A. Erives, M. Markstein, M. Levine, Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo, *Cell* **111**, 687–701 (2002).

63. F. Biemar *et al.*, Comprehensive identification of Drosophila dorsal-ventral patterning genes using a whole-genome tiling array, *Proc Natl Acad Sci U S A* **103**, 12763–12768 (2006).

64. J.-W. Hong, D. A. Hendrix, M. S. Levine, Shadow enhancers as a source of evolutionary novelty, *Science* **321**, 1314 (2008).

65. S. D. Hanes, G. Riddihough, D. Ish-Horowicz, R. Brent, Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen, *Molecular and Cellular Biology* **14**, 3364–3375 (1994).

66. H.-L. Liang *et al.*, The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila, *Nature* **456**, 400–403 (2008).

67. J. R. ten Bosch, J. A. Benavides, T. W. Cline, The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription, *Development* **133**, 1967–1977 (2006).

68. M. M. Harrison, M. R. Botchan, T. W. Cline, Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes, *Developmental Biology* **345**, 248–255 (2010).

69. C.-Y. Nien *et al.*, Temporal coordination of gene networks by Zelda in the early Drosophila embryo, *PLoS Genetics* **7**, e1002339 (2011).

70. M. M. Harrison, X.-Y. Li, T. Kaplan, M. R. Botchan, M. B. Eisen, Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition, *PLoS Genetics* **7**, e1002266 (2011).

71. R. Satija, R. K. Bradley, The TAGteam motif facilitates binding of 21 sequence-

specific transcription factors in the Drosophila embryo, *Genome Res.* **22**, 656–665 (2012).

72. U. Mayer, C. Nüsslein-Volhard, A group of genes required for pattern formation in the ventral ectoderm of the Drosophila embryo, *Genes & Development* **2**, 1496–1511 (1988).

73. M. Freeman, Rhomboids: 7 years of a new protease family, *Semin. Cell Dev. Biol.* **20**, 231–239 (2009).

74. E. Bier, L. Y. Jan, Y. N. Jan, rhomboid, a gene required for dorsoventral axis establishment and peripheral nervous system development in Drosophila melanogaster, *Genes & Development* **4**, 190–203 (1990).

75. P. P. Szymanski, M. M. Levine, Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo, *EMBO J.* **14**, 2229–2238 (1995).

76. L. Bintu *et al.*, Transcriptional regulation by the numbers: models, *Current Opinion in Genetics & Development* **15**, 116–124 (2005).

77. A. Ay, D. N. Arnosti, Mathematical modeling of gene expression: a guide for the perplexed biologist, *Critical Reviews in Biochemistry and Molecular Biology* **46**, 137–151 (2011).

78. G. K. Ackers, A. D. Johnson, M. A. Shea, Quantitative model for gene regulation by lambda phage repressor, *Proc Natl Acad Sci U S A* **79**, 1129–1133 (1982).

79. I. Mogno *et al.*, TATA is a modular component of synthetic promoters, **20**, 1391–1397 (2010).

80. R. P. R. Zinzen, K. K. Senger, M. M. Levine, D. D. Papatsenko, Computational models for neurogenic gene expression in the Drosophila embryo, *Current Biology* **16**, 1358–1365 (2006).

81. H. Janssens *et al.*, Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene, *Nat Genet* **38**, 1159–1165 (2006).

82. W. D. Fakhouri *et al.*, Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo, *Mol. Syst. Biol.* **6**, 341 (2010).

83. E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, U. Gaul, Predicting expression patterns from regulatory sequence in Drosophila segmentation, *Nature* **451**, 535–540 (2008).

84. M. D. Schroeder *et al.*, Transcriptional Control in the Segmentation Gene Network of Drosophila, *PLoS Biol* **2**, e271 (2004).

CHAPTER II

Optimization of reporter gene architecture for quantitative measurements of gene expression in the *Drosophila* embryo[1]

[1]The work described in this chapter was published as the following manuscript: Rupinder Sayal, Seuk-Min Ryu, and David N. Arnosti (2011). Optimization of reporter gene architecture for quantitative measurements of gene expression in the *Drosophila* embryo. Fly (Austin), 5(1):47-52.

Abstract

Quantitative assessment of gene regulation is critical for mathematical modeling of transcriptional systems for systems biology efforts. Enhancers, also termed *cis*-regulatory modules (CRMs), are the primary mediators of transcriptional regulation in higher eukaryotes; transcription factors binding to CRMs dictate the likelihood and frequency of promoter activation. To provide a suitable platform for in-depth CRM analysis, we adapted a targeted integration vector to compare action of basal promoters with diverse combination of TATA, Inr and DPE motifs, as well as a set of 3'-UTRs representative of those used in different reporter vectors. This “Honda” series of reporter gene vectors was activated by a regulatory element binding Dorsal and Twist activators suitable for transcription in the early Drosophila embryo. The diverse promoters functioned in a similar manner with minor quantitative differences, consistent with a lack of enhancer-promoter specificity. Constructs bearing SV40 3'-UTR sequences appeared to produce somewhat higher levels of mRNA.

Confocal laser scanning microscopy revealed that the mRNA distribution produced by these constructs was punctate; this pattern appears to be dependent on 5'-UTR sequences, as an optimized vector including an alternate 5'- UTR produced a more even distribution, which may be preferable for quantitative imaging. This set of Honda vectors contains convenient sites for modification of basal promoter, 3'-UTR, and enhancer, and will be useful for analysis of CRMs and quantitative studies of gene expression.

Introduction

Systems biology studies have provided rich new data sets pertinent to understanding transcriptional regulation, which is critical for tissue differentiation and physiological adaptation. Non-coding DNA elements responsible for regulation of gene expression in eukaryotes include binding sites for sequence-specific transcription factors and for core transcriptional machinery. Phylogenetic comparisons, *in vivo* binding information about different transcription factors, and global transcriptome data have revealed a plethora of putative *cis*-regulatory elements. Experimental validation of these elements has lagged behind their discovery, however.

Drosophila melanogaster is a well-established model system of metazoan development, differentiation and disease. Functional characterization of enhancers and their evolution in this organ- ism can provide insights into developmental regulatory network architecture. In *Drosophila*, P-element-mediated transgenesis has been used extensively to test the activities of *cis*-regulatory elements driving the expression of reporter genes, which are then often monitored by *in situ* hybridization with RNA probes or by expression of fluorescent proteins. More recently, φC31-mediated targeted integration has been adopted to limit the variability due to position effects, which complicate interpretation of quantitative expression levels (*1*). Reporter gene analysis should, but generally does not, also consider effects that may be caused by the 5' and 3'-UTR, as well as the basal promoter, the 100 bp flanking the transcriptional start site.

The RNA polymerase II core promoter is sufficient for basal transcription in vitro and directs initiation in vivo when driven by adjoining cis-regulatory sequences. In the past few decades, various core promoter sequence motifs have been discovered that

direct interaction with components of basal transcription machinery (*2*). Diverse core promoter structure can influence the selectivity and operation of gene transcription by interacting with different components of core transcriptional machinery, providing enhancer specificity and permitting pre-loading of basal transcription machinery prior to firing of a gene (*3*). Although enhancers and basal promoters can often exhibit compatibility, in certain cases, enhancers can be highly selective towards the type of promoter they interact with (*4*, *5*).

5'- and 3'-UTRs of an mRNA have important roles in regulation of its localization and stability, which can be dictated by sequence motifs important for interaction with miRNAs and RNA-binding proteins. In quantitative studies of gene expression, mRNA stability influences the pool size of the transcript and the response of the system to dynamic changes. A relatively unstable mRNA may be less easily detected but may more readily reflect temporal shifts in gene activity. The localization of mRNA can be also crucial for quantitative studies that determine the number of transcripts associated with expression originating in a particular nucleus, as in the syncytial environment of *Drosophila* blastoderm embryo. Recent studies have demonstrated a wide variety of mRNA localization patterns, which can influence developmental patterning, stem cell fate and cell division (*6*).

In developing platforms useful for quantitative analysis of gene expression, there has been little focus on systematic exploration of promoter and UTR effects. We have developed a suite of φC31-based reporter gene vectors designed to address this issue. For the set of commonly used basal promoters and 3'-UTRs, our direct comparison reveals modest effects of basal promoter structure and 3'-UTR sequence on gene

expression measured in the blastoderm embryo. These constructs provide a flexible platform suitable for quantitative analysis of *Drosophila* enhancers.

Results

Design of pHonda vectors. To establish a platform suitable for quantitative assessment of enhancer function, we modified the pattB vector to incorporate a cloning site for an enhancer, a 340 bp neutral spacer sequence (previously confirmed to lack transcriptional activity (*7*)), a set of interchangeable basal promoter regions extending 50 bp 5' and 3' of the transcriptional start site, and a *lac*Z-based reporter gene with various 3'-UTR sequences (Fig. 2.1). The spacer between promoter and enhancer reduces the possibility of direct effects by short-range repressors and tests the distally-acting sequences in a more natural configuration. To facilitate insertion of different CRMs, we created a multiple cloning site that contains unique restriction sites for *Age*I, *Asc*I, *Avr*II, *Pac*I and *Fse*I.

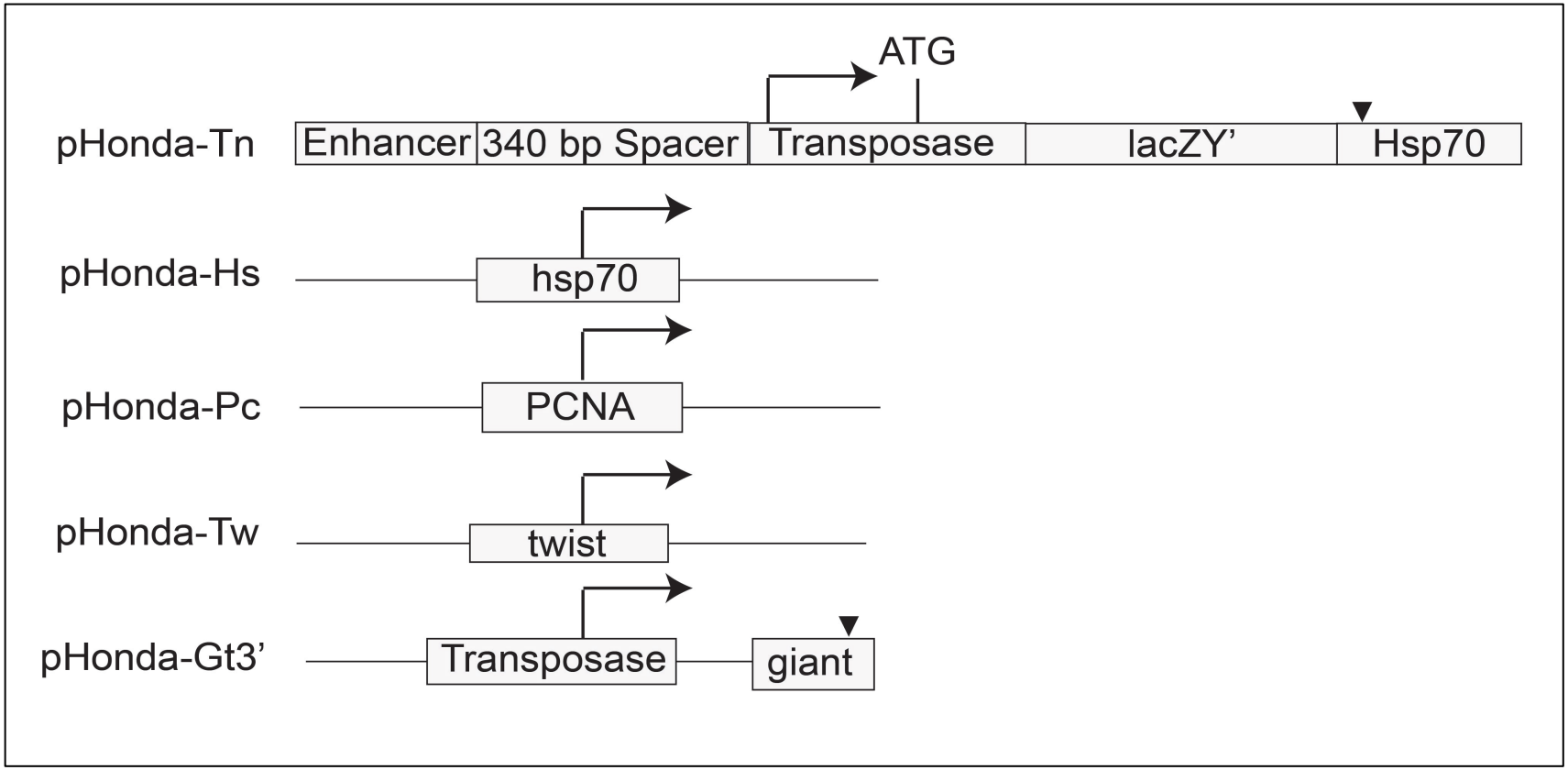
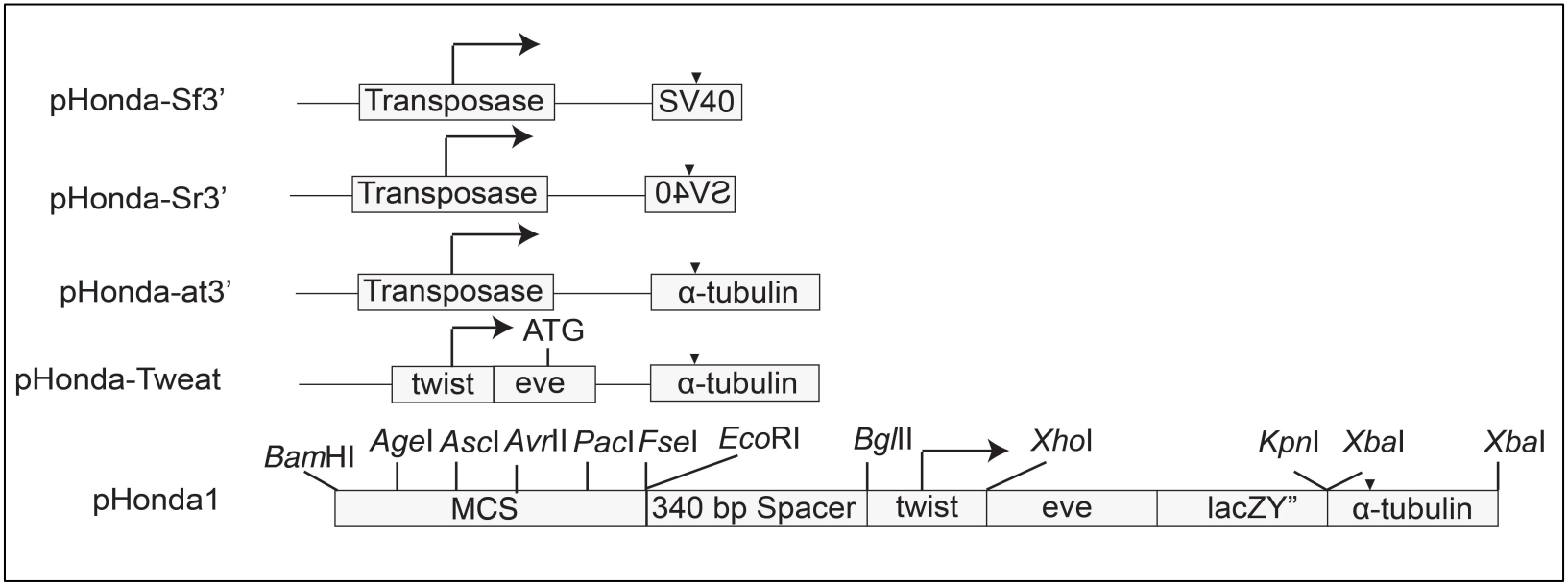


Figure 2.1. A schematic illustration of constructs. All constructs (except phonda1) are driven by a synthetic regulatory element containing Dorsal, Twist and Knirps binding sites. Constructs have a 340 bp spacer derived from the *kni* ORF between enhancer and transcription unit. pHonda-Tn has the transposase promoter 5'-UTR, *lac*Z coding region and a small segment of *lac*Y and *hsp*70 3'-UTR. pHonda-Hs is identical except for the *hsp*70 basal promoter. Similarly, pHonda-Pc and pHonda-Tw have *PCNA* and *twist* promoters respectively. pHonda-Gt3', pHonda-Sf3', pHonda-Sr3' and pHonda-

Figure 2.1 cont'd

Ta3' have *giant*, SV40 (both orientations) and α-*tubulin* 3'-UTRs respectively. The black triangle indicates the polyadenylation site within the 3'-UTR. pHonda-Tweat has the *twist* core promoter, *eve* 5'-UTR, *lac*Z and a portion of *lac*Y and α-*tubulin* 3'-UTR. pHonda1 has a multiple-cloning site for insertion of novel CRMs instead of the synthetic regulatory element used in this study.



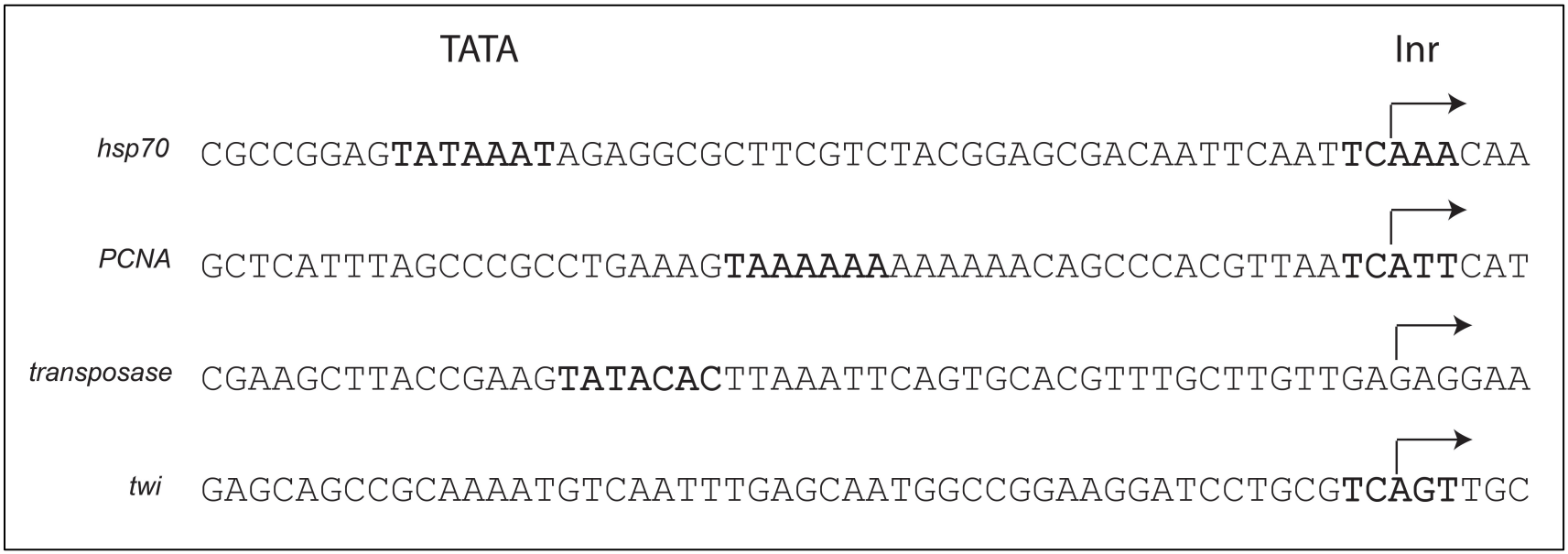
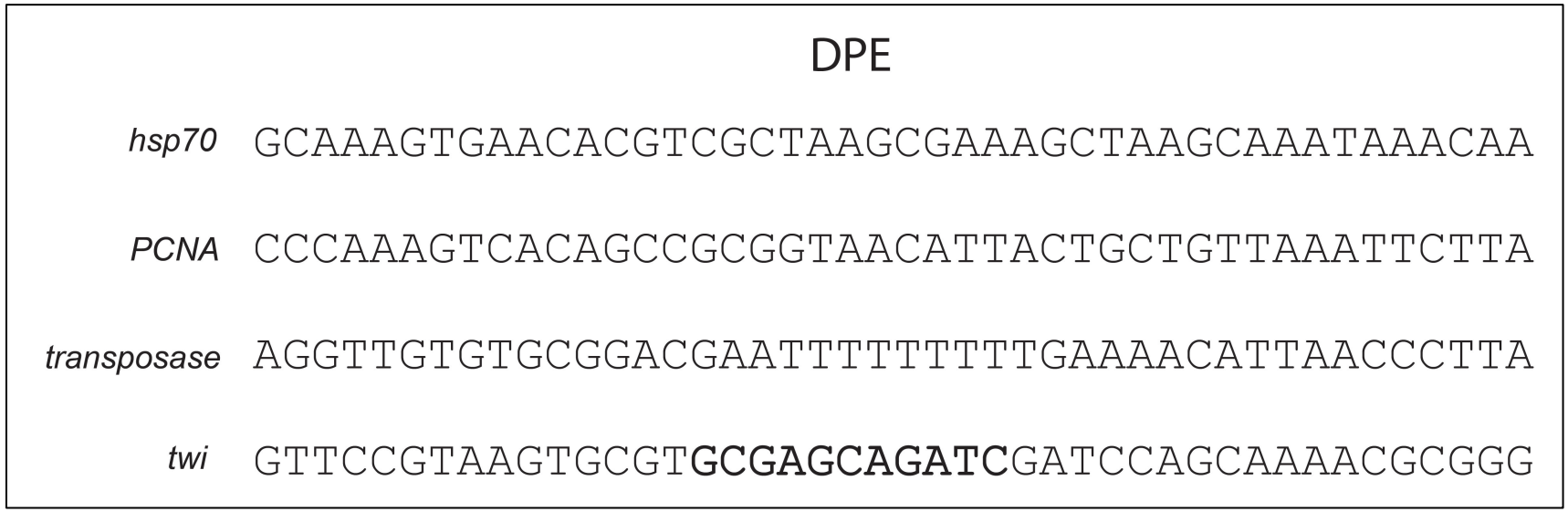


Figure 2.2. Promoter sequences tested in this study, with core promoter motifs highlighted in bold—TATA box, Initiator and DPE. Core promoter motifs were scored and identified using MAST.

Figure 2.2 cont'd



Different promoters and 3'-UTRs direct equivalent amounts of mRNA. To test the effects of promoter architecture on reporter gene expression, we assayed four different promoters with divergent structure that represent a part of the spectrum of core promoter motifs found in the *Drosophila* genome (Fig. 2.2). The transposase promoter has been used extensively in enhancer studies and also in quantitative studies of gene expression (*8*, *9*); this element contains a TATA box. The widely used *hsp*70 promoter and the PCNA promoter contain a TATA box and Initiator (Inr) motif involved in TFIID interactions (*10*). The *twi* promoter contains an Inr motif and a DPE motif (*11*). Two other motifs sometimes associated with *Drosophila* core promoters, Pause Button (PB) (*12*) and Motif Ten Element (MTE), were not identified in these promoters. The promoters were joined to a *lacZ* reporter gene, and the expression was driven in ventral regions of the *Drosophila* embryo by Dorsal and Twist activators. Embryos containing a single copy of each transgene were collected, fixed and hybridized in parallel to detect *lacZ* mRNA. Semi-quantitative alkaline phosphatase staining revealed that the four promoters appeared to provide similar output in response to Dorsal and Twist activation, and that the Knirps repressor sites mediated similar repression in all four constructs (Fig. 2.3).

To investigate whether different 3'-UTRs have a significant effect on reporter gene expression, we tested five different 3'-UTRs, including some commonly used reporters: *hsp*70, *giant* (*gt*), *α-tubulin* (*αTub84B*) and both orientations of the SV40 UTR region, which is used for transcripts originating from either directions in the viral genome (Fig. 2.1). We found that constructs bearing either orientation of the SV40 3'-UTR tended to exhibit somewhat elevated transcript levels, while constructs with the *hsp70*,

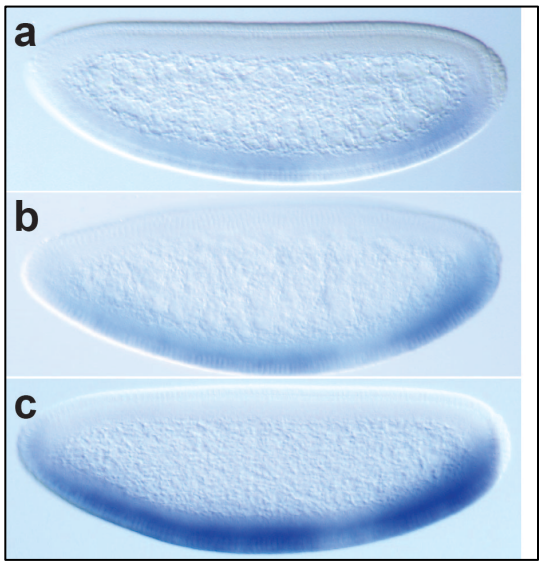


Figure 2.3. Expression levels of pHonda transgenes assayed in *Drosophila* blastoderm embryos by in situ hybridization. Ventral expression is directed by Dorsal and Twist activators. Gap in posterior region reflects repression by Knirps. (A–C) show representative light, medium and dark staining observed from in situ mRNA hybridization against antisense digU-labeled RNA probe. The four basal promoters tested here gave similar results, with a possible slightly higher expression with transposase promoter. The SV40 forward and reverse constructs (phonda-Sf3' and phonda-Sr3') appeared to be associated with more robust steady state levels of mRNA.

Table 3a Staining counts for constructs

Construct	Light	Medium	Dark	n
pHonda-Hs	92	8	0	66
pHonda-Pc	83	17	0	30
pHonda-Tn	75	19	6	85
pHonda-Tw	91	9	0	130
pHonda-Gt3'	84	16	0	106
pHonda-Sf3'	69	27	4	55
pHonda-Sr3'	55	34	10	29
pHonda-at3'	73	19	9	81

Table 3b Effect of copy number on staining intensity

Sample	Light	Medium	Dark	n
pHonda-at3' (1 copy)	83	12	5	60
pHonda-at3' (2 copies)	10	23	67	252

Table 3a shows the relative expression levels of different reporter constructs, expressed as percentages of embryos with light, medium or dark staining.

Table 3b shows the percentages of embryos with the indicated phenotypes for heterozygous or homozygous embryos with the phonda-at3' construct. The dynamic range of this assay can clearly distinguish twofold effects, and the differences among constructs observed in 3a appear to represent less than twofold effects. Embryos in each table were stained in parallel (n = number of stained stage 5 embryos).

α-tubulin and *giant* 3'-UTRs were expressed at similar lower levels (Fig. 2.3). To relate the different levels of staining observed here to quantitative differences in mRNA levels, we stained heterozygous and homozygous lines for one construct to measure the differences produced by doubling gene dosage (Fig. 2.3). We found that the majority of stainings shifted from the lightest category to the darkest category upon doubling of gene copy number, indicating that this assay is sensitive to twofold changes in mRNA. We conclude that the differences observed due to alternative 3'UTR sequences, which were less dramatic than these effects, represent less than twofold differences in gene expression. For most gene expression studies, such differences would still be tolerable, provided identical elements are incorporated in constructs that are to be compared with each other.

5'-UTRs affect mRNA localization. Using confocal laser scanning microscopy in previous studies, we had noticed that a construct similar to pHonda-Tn bearing the transposase promoter and 5'-UTR and *hsp*70 3'-UTR produced a punctate distribution of *lac*Z mRNA(*13*). This effect is not apparent with other *lac*Z reporters utilized in embryonic studies(*14*), therefore we tested whether promoter or 3'-UTR sequences had an influence. Confocal microscopy of the pHonda variants described above revealed a similar punctate distribution (Fig. 2.4; data not shown). Therefore, we tested an alternative 5' UTR sequence derived from the *eve* gene, which is present in *eve-lac*Z fusion reporter genes (*15*). The transcripts produced by this gene, pHonda-Tweat, were more evenly distributed than those of pHonda-at3' bearing the transposase 5'-UTR (Fig. 2.4). Average expression levels were very similar, indicating that the alternative 5'-UTR structure does not affect expression levels, but affects localization. This reporter gene

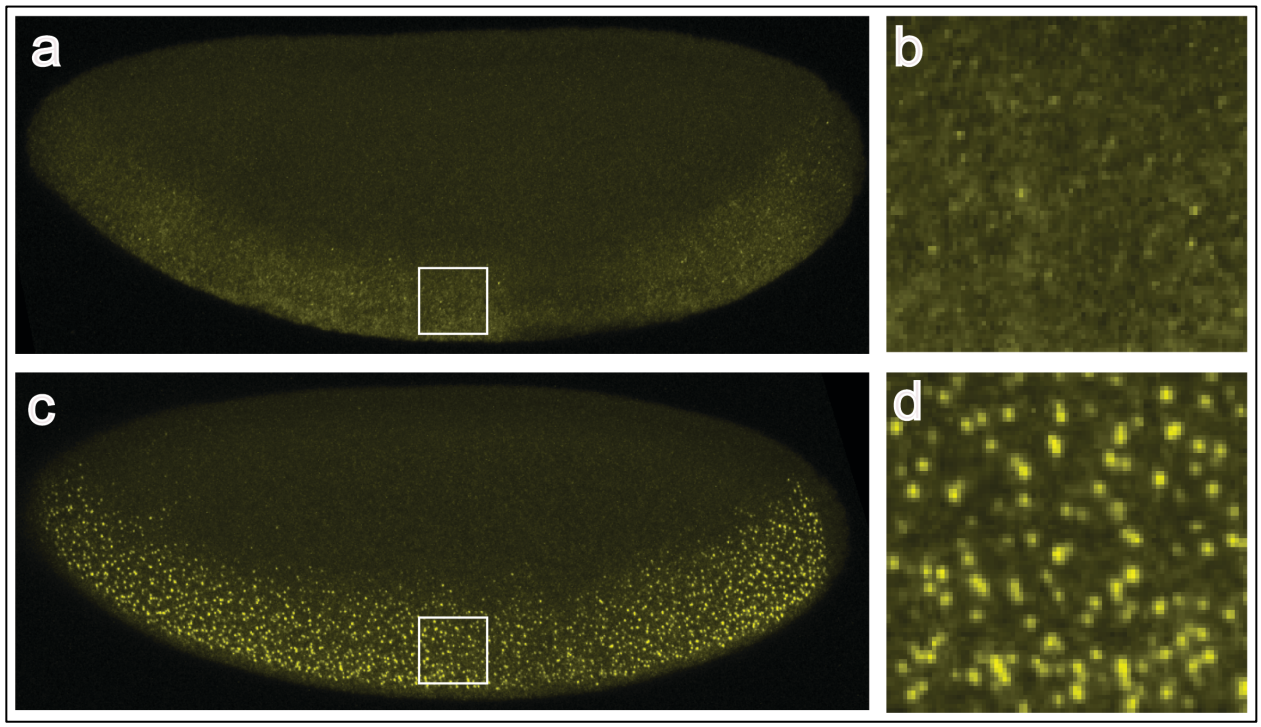


Figure 2.4. Distinct patterns of mRNA localization directed by *transposase* and *eve* 5'-UTRs. (a) Confocal laser scanning microscopy image of a Drosophila embryo that is *trans*genic for pHonda-Tweat and stained for *lacZ* (eve 5'-UTR) and (c) pHonda-at3' (*transposase* 5'-UTR). (b and d) Magnified view from (a and c) showing diffuse mRNA localization pattern for pHonda-Tweat and punctate pattern for phonda-at3'. The background levels were the same among these embryos. In areas of staining, average intensity levels were similar. (Mean ± SD for pHonda-Tweat-62 ± 7, n = 6; pHonda-at3'-67 ± 11, n = 10).

may permit more facile quantitative analysis of relative mRNA levels, with less need for averaging (*13*). We produced a derivative of this construct that contains a set of unique restriction sites in place of the upstream enhancer, to permit analysis of additional CRMs (pHonda1, Fig. 2.1).

Discussion

Transcriptional reporter genes are important tools for testing putative enhancers to deepen understanding of developmental regulatory networks. In this study, we found that basal promoter structure had little effect on reporter gene output in response to Dorsal and Twist activators, indicating permissive enhancer-promoter interactions. Other enhancers may demonstrate more restrictive interactions. Thus, for investigation of cis elements of a particular gene, researchers may wish to insert the cognate core promoter region into the pHonda1 vector.

3'-UTRs have been demonstrated to be important for determining RNA stability. In this study, we determined that SV40 sequences appeared to support somewhat more robust levels of mRNA, judging by *in situ* mRNA hybridization. Less stabilizing 3'-UTRs may permit more accurate estimation of dynamic synthesis levels, as in pHonda1. Researchers interested in studying different aspects of mRNA stability or localization may wish to insert different 3'-UTRs.

In addition to overall levels of mRNA produced by the transgene, the mRNA localization can also affect quantitative analysis of gene expression. Our study found that the 5'-UTR of lacZ reporter, derived from a widely used P-element vector, caused punctate accumulation of mRNAs, which may complicate quantitative analysis by obscuring fine features of patterns and necessitating smoothening algorithms. pHonda1

employs an alternative leader sequence that is optimized for quantitative analysis of embryonic reporters.

Previously used P-element-based vectors integrate randomly, inducing positional effects that complicate the analysis and comparison of different enhancers. Our pHonda1 vector is based on φC31-based transgenesis system to allow testing single lines of reporter gene constructs, making comparisons easier and faster. These features make our vector suite an ideal platform for testing and analyzing enhancers in *Drosophila*.

Materials and Methods

Vector Construction. For comparison of promoter and UTR elements, we employed a common 165 bp enhancer containing 4 Dorsal, 4 Twist and 2 Knirps binding sites (*9*), which was synthesized using overlapping oligos and cloned into *Bam*HI and *Eco*RI sites of pattB vector (*1*). The enhancer sequence is: 5'-ACC GGT GGG AAA ACC CAA AAT CGA GGG ATT TTC CCA TCT AGA **CAT ATG** CTC AA**C ATA TG**G GAT CC**C TGA TCT AGT TT**G TAC TAG ACA T**CT GAT CTA GTT T**GG ATC C**CA TAT G**TT GAG **CAT ATG** TCT AGA GGG ATT TTC CCA AAT CGA GGG AAA ACC CAA GGC CGG CC-3'. (The Dorsal sites are underlined, Twist sites are bold, and Knirps sites are underlined and bold). A 340 bp neutral spacer derived from the *kni* open reading frame was amplified from *D. melanogaster* genomic DNA and cloned into *Eco*RI and *Bgl*II sites, with following primers: 5'-GCG AAT TCA ACC GCT TTA GTC CCG CCA G-3' and 5'-AGC CAG ATC TTG TGC ACG GAG CTC CGC GAG-3'. Basal promoters tested in this study comprised of 100 nt regions (-50 to +50) synthesized from overlapping oligos and cloned into *Bgl*II and *Xho*I sites. The transposase 5'-UTR, *lac*ZY transcription unit

and *hsp*70 3'-UTR fragment were amplified from C4PLZ (*8*) and cloned into *Xho*I and *Kpn*I sites using primers 5'-CGC TCG AGC GTG GAA TAA AAA AAA ATG AAA TAT TGC-3' and 5'-GGC GGT ACC GAT CTA AAC GAG TTT TTA AGC-3'. The resulting transcription unit contains a 5'-UTR consisting of 50 nucleotides derived from the basal promoter, 16 nucleotides derived from 5'-UTR and 440 nucleotides from *transposase* coding region. For genes pHonda-e and pHonda1, the transposase segment was replaced with an *eve* 5'-UTR and *lac*ZY transcription unit: This fragment contains the *eve* 5'-UTR starting from +50 position and continuing until codon 22 of *eve* open reading frame, fused with codon 7 of *lac*Z and continuing with a portion of *lac*Y. The fragment was amplified from pCasper-rhoNEE-eve-lacZ (a gift from Albert Erives (*14*)) and cloned into *Xho*I and *Kpn*I sites using primers 5'-CAG GCG CTC GAG TTA ATA TCC TCT GAA TAA GCC-3' and 5'-GTG GCG GGT ACC GGG CCT AGA GCT TGC CGA GTT TGT C-3'. To generate pHonda1, the pattB vector was modified by insertion of an enhancer multiple cloning site created by two DNA oligos containing restriction sites for enzymes *Age*I, *Asc*I, *Avr*II, *Pac*I and *Fse*I (in that order) cloned into the *Bam*HI and *Eco*RI sites of the pattB vector. The oligos used were: 5'-GAT CCA TGA CCG GTG ATG GCG CGC CTG CCC TAG GGC ATT AAT TAA TGC GGC CGG CCG AG-3' and 5'-AAT TCC TCG GCC GGC CGC ATT AAT TAA TGC CCT AGG GCA GGC GCG CCA TCA CCG GTC ATG-3'. All 3'-UTRs were cloned into the *Xba*I site. For the *giant* (*gt*) 3'-UTR, a 383 bp fragment containing the 324 bp long UTR was amplified from *D. melanogaster* genomic DNA using oligos 5'-GAT TCT AGA AGG TCC ACT CCT CTC TTG AT-3' and 5'-GTA TCT AGA AAA TTA CCA GGC GAA CAG GA-3' and the 285 bp *αTub84B* 3'-UTR contained within an 800 bp fragment was similarly amplified using

oligos 5'-GCG TCT AGA AGT ACT AAG CGT CAC GCC AC-3' and 5'-GCG TCT AGA TTG CCT AAT TGT TTC AGA TTT ATG GGG-3'. For the SV40 3'-UTR, a 202 bp 3'-UTR for SV40 late gene was amplified from pRL-CMV (Promega) using oligos 5'-GAT TCT AGA GAT GAG TTT GGA CAA ACC AC-3' and 5'-GTAT CTA GAT ACC ACA TTT GTA GAG GTT TTA C-3'. The forward direction represents the strand used for SV40 late transcripts. The reporter gene constructs made in this study are available upon request.

Transgenesis and fly stocks. φC31-mediated transgenesis was carried out in-house and by Rainbow Transgenic Flies, Inc., using the 51D cytological location (*1*). The in situ mRNA hybridization for *lacZ* was performed as previously described in reference 15. Immunofluorescent in situ hybridization and confocal laser scanning microscopy was done as previously described and was done in parallel (*9*). Heterozygous embryos with only one copy of transgene were collected for in situ hybridization. Image analysis was done with ImageJ (*16*).

Motif analysis. The four promoter sequences were searched using log-odds matrices for TATA, Inr, DPE and MTE motifs (*17*) and analyzed by MAST (*18*). The Pause Button motif was searched for using its IUPAC notation, KCGRWCG (*12*).

Acknowledgements

We would like to thank Konrad Basler (Institute of Molecular Life Sciences University of Zurich, Switzerland) for the pattB vector, Albert Erives (Dartmouth College, NH USA) for pCasper-rhoNEE-lacZ construct, Steve Small (New York University, NY USA) for the information about the sequence of *αTub84B* 3'-UTR, Melinda Frame of CAM (MSU) for

help with confocal microscopy and members of Arnosti lab for helpful discussions. This study was supported by grant NIH GM56976 to David N. Arnosti.

REFERENCES

REFERENCES

1. J. Bischof, R. K. Maeda, M. Hediger, F. Karch, K. Basler, An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases, *Proceedings of the National Academy of Sciences of the United States of America* **104**, 3312–3317 (2007).

2. T. Juven-Gershon, J. T. Kadonaga, Regulation of gene expression via the core promoter and the basal transcriptional machinery, *Developmental Biology* **339**, 225–229 (2010).

3. J.-Y. Hsu *et al.*, TBP, Mot1, and NC2 establish a regulatory circuit that controls DPE-dependent versus TATA-dependent transcription, *Genes & Development* **22**, 2353–2358 (2008).

4. J. E. Butler, J. T. Kadonaga, Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs, *Genes & Development* **15**, 2515–2519 (2001).

5. C. Merli, D. E. Bergstrom, J. A. Cygan, R. K. Blackman, Promoter specificity mediates the independent regulation of neighboring genes, *Genes & Development* **10**, 1260–1270 (1996).

6. E. Lécuyer *et al.*, Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function, *Cell* **131**, 174–187 (2007).

7. M. M. Kulkarni, D. N. Arnosti, cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster, *Molecular and Cellular Biology* **25**, 3411–3420 (2005).

8. K. A. Wharton, S. T. Crews, CNS midline enhancers of the Drosophila slit and Toll genes, *Mech. Dev.* **40**, 141–154 (1993).

9. W. D. Fakhouri *et al.*, Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo, *Mol. Syst. Biol.* **6**, 341 (2010).

10. T. W. Burke, J. T. Kadonaga, The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila, *Genes & Development* **11**, 3020–3031 (1997).

11. B. Thisse, C. Stoetzel, C. Gorostiza-Thisse, F. Perrin-Schmitt, Sequence of the twist gene and nuclear localization of its protein in endomesodermal cells of early Drosophila embryos, *EMBO J.* **7**, 2175–2183 (1988).

12. D. A. Hendrix, J.-W. Hong, J. Zeitlinger, D. S. Rokhsar, M. S. Levine, Promoter elements associated with RNA Pol II stalling in the Drosophila embryo, *Proc Natl Acad Sci U S A* **105**, 7762–7767 (2008).

13. A. Ay, W. D. Fakhouri, C. Chiu, D. N. Arnosti, Image processing and analysis for quantifying gene expression from early Drosophila embryos, *Tissue Eng Part A* **14**,

1517–1526 (2008).

14. J. Crocker, Y. Tamori, A. Erives, Evolution acts on enhancer organization to fine-tune gradient threshold readouts, *PLoS Biol* **6**, e263 (2009).

15. S. Small, A. Blair, M. Levine, Regulation of even-skipped stripe 2 in the Drosophila embryo, *EMBO J.* **11**, 4047–4057 (1992).

16. M. Abràmoff, P. Magalhães, Image processing with ImageJ, *Biophotonics International* **11**, 36–42 (2004).

17. U. Ohler, G.-C. Liao, H. Niemann, G. M. Rubin, Computational analysis of core promoters in the Drosophila genome, *Genome Biol* **3**, RESEARCH0087 (2002).

18. T. L. Bailey, M. Gribskov, Combining evidence using p-values: application to sequence homology searches, *Bioinformatics* **14**, 48–54 (1998).

Chapter III

Deep perturbation analysis and transcriptional modeling to predict an embryonic regulon

Abstract

A major genomics era challenge is comprehensive interpretation of transcriptional regulatory elements. Experimental studies provide a picture of the transcriptional “state” of cells, yet systems-level understanding demands quantitative tools that build on these data for comprehensive insights into the significance of genetic variation affecting transcriptional elements. Thermodynamic models can quantitatively predict DNA-based transcriptional information, but previous efforts have relied on a limited number of models applied to sparse data sets. Here, we describe a novel, deep perturbation analysis targeted at understanding the early embryonic regulon of the Drosophila NF-κB homolog Dorsal. We investigate 120 models to characterize transcription factor cooperativity and quenching on the *rhomboid* enhancer and related elements. Application of these models to early embryonic enhancers of the regulon shows that analysis based on deep perturbation of a representative cis-regulatory element is capable of predicting essential patterning information for a group of 47 genes that share a common transcriptional “grammar”, opening a door to more comprehensive analysis of developmental systems.

Introduction

Developmentally expressed genes in metazoans are commonly regulated by diverse cis-regulatory elements, including distally-acting sequences termed enhancers. Enhancers, which typically range in size from 100 bp-1kb and usually feature binding sites for multiple transcription factors, play crucial roles in development, differentiation and emergence of novel morphological features in evolution (*1*-*3*). Recently, a number of studies have characterized several cases of apparently redundant 'shadow' or 'distributed' enhancers which may be required to buffer transcriptional output of genes against genetic and environmental perturbations (*4*-*7*).

Despite almost three decades of progress on discovery and characterization of enhancers, surprisingly little is known about the internal structural organization of binding sites within the elements. Relative positions of binding sites can be tightly constrained in some cases, as in enhanceosome-like elements, which permit no change in spacing without catastrophic effects on function (*8*, *9*). In other, less tightly-constrained genetic elements, sometimes termed "billboard enhancers" there is evidence for preferred spacing between transcription factor sites, but many functionally conserved elements exhibit a large degree of evolutionary variation, suggesting that the influences of such relationships are subtle (*10*-*13*). A better understanding of the internal enhancer "grammar" of these *cis*-regulatory elements would permit researchers to better understand the significance of genetic variation that is observed within and between species. In a number of cases, additions of binding sites or spacing between transcription factor motifs have been linked to evolution of morphological innovations or disease-related phenotypes.

Transcriptional regulatory elements have been traditionally studied by use of reporter gene assays, which can reveal the function of individual binding sites for transcription factors (*14*). Recently, genome-wide experimental analysis of transcriptional regulation has moved biology closer to systems-level understanding of gene expression, and the many roles that this process plays in development and disease. Enhancers and related regulatory sequences can be identified from clusters of putative binding sites or actual in vivo occupancy by transcription factors, transcriptional cofactors including CBP/p300, and distinctive chromatin modifications (*15*-*17*). Despite encouraging progress in this field, the genome-wide studies are invariably limited to providing a "snapshot" of a cell's or organism's transcriptional status, only a small fraction of the total possible configurations that a cell might undergo during development or physiological adaptation. In addition, although comprehensive, the functional relevance of many of the features measured is unknown; in vivo binding events or chromatin modifications may be of no physiological relevance. Thus there is a compelling need to develop systematic, quantitative tools that can provide reliable predictive indications of the range of possible activities of particular DNA regulatory elements, and complement the type of information provided by ChIP-seq and other methods.

To meet this challenge, mathematical "thermodynamic" or fractional occupancy models have been used to provide a framework for understanding how transcription factors interacting with specific DNA sequences regulate gene expression (*18*). These models employ tools from statistical physics to model gene activity as a function of protein-DNA and protein-protein interactions. The probability of gene expression under

a steady-assumption is functionally related to the population of the enhancer region in “active” vs. “inactive” states, which are influenced by the levels of regulatory factors, the presence of binding motifs in the enhancer, and inferred functional interactions amongst the transcription factors, cofactors, and basal machinery (*19*-*23*). A variety of new developments supports the wider application of these models. Comprehensive analysis of quantitative temporal and spatial gene expression is becoming widely available through community resources (*24*-*26*). Our knowledge of specificity of transcription factor binding preferences has also expanded through application of high-throughput approaches. In addition, genome-wide binding maps of numerous transcription factors in cells and in whole animals have been made available (*27*-*31*). These studies have pointed to wide-spread binding of transcription factors to thousands of genomic regions in a “quantitative continuum”, emphasizing the importance of utilizing quantitative models to distinguish functional and non-functional binding, as all binding events may not be functional (*32*).

Previous efforts at thermodynamic modeling in eukaryotic systems have demonstrated that this method is capable of fitting diverse types of data, and generating predictions that match known transcriptional responses, at least at a qualitative level (*22*, *23*, *33*). Two major limitations have blunted the impact of such efforts so far; first, for high-resolution analysis that informs us of the transcriptional “grammar” that can influence enhancer output, thermodynamic models require quantitative information about binding preferences of transcription factors. High quality perturbation datasets are also essential to allow the modeler to test the possible functional relationships between transcription factors. A recent analysis of synthetic yeast promoters provides a good

example of such an effort; no comparable effort has been reported in metazoan systems, whose transcriptional elements feature activities not found in yeast (*34*). Most research efforts that have based their modeling on experimentally measured transcriptional enhancers have relied on existing datasets of the expression patterns of multifarious regulatory elements that bind to a diverse set of transcription factors. The composition of the elements in question is heterogeneous, leading to a diminished power to infer specific features that contribute quantitatively to the output of specific enhancers.

Second, thermodynamic models offer the possibility of capturing with the help of specific parameters biochemical activities that underlie the subtle context-dependence of binding sites; these parameters can include the cooperativity seen between activators, or repressor-activator antagonism. Studies tested on Drosophila transcriptional elements, which represent the majority of thermodynamic modeling research in metazoans, have not explored this aspect extensively, however. For example, the highly distance-sensitive “quenching” of transcriptional activators by short-range repressors has been represented by step or monotonically declining functions, but more recent research indicates that this function is a complex, non-monotonic one (*20*). To address both of these limitations, here we describe a novel in-depth perturbation analysis that takes advantage of the quantitative setting of the Drosophila blastoderm to generate a very high-quality data set that is then in turn subject to a comprehensive set of thermodynamic models. The results of the modeling effort are validated not only on the modeled data set, but tested in a realistic manner on the highly heterogeneous sequences that are coordinately controlled by the Drosophila homolog

of NF-κB, the Dorsal protein. Our study suggests that for systems in which high-quality data can be generated, such mathematical approaches can provide novel system-wide predictive power, with the possibility to expand our knowledge of genetic variation on a global scale.

Materials and Methods

Reporter gene constructs

The 318-bp *rhomboid* neurectodermal enhancers were cloned into *Age*I and *Fse*I restriction sites of the pHonda1 pattB-based targeted integration vector (*35*, *36*). Enhancers were assembled from 40-60 bp overlapping synthetic 5'-phosphorylated oligonucleotides with 10 bp overhangs, which were annealed, and then ligated into pHonda1. The footprinted binding sites for Dorsal, Twist and Snail, as well as two predicted E-box motifs thought to be bound by bHLH factors, were mutated as follows using sequences previously shown to affect *rho* enhancer activity (*35*, *37*):

Dorsal1 - GGGAAAAACAC to TTTAAAAACAC

Dorsal2 - CGGAATTTCCT to CGTCAGTTAAT

Dorsal3 - GGGAAATTCCC to TCTAGATTATC

Dorsal4 - GGGAAAGGCCA to AGGCCTGGTCA

Twist1 - CGCATATGTT to ACGCGTTGTT*

Twist2 - AGCACATGTT to ACGCGTTGTT

Snail1 – CAACTTGCGG to CAGAGCTCGG

Snail2 - CACCTTGCTG to CAGGAGCTTG*

Snail3 – CCACTTGCGCT to CCGCCGGCGT*

Snail4 – GCACATGTTT to GCATATGTTT

bHLH1 - CATTTG to TGATTC*

bHLH2 - CAAGTG to TAGCGA*

(*novel mutations developed for this study)

The bHLH1 and bHLH2 sites were mutated simultaneously. The mutations are predicted to reduce the binding score for each transcription factor to near background values. Additional wild-type enhancers for other genes were created by PCR amplification from genomic DNA or by assembly using oligonucleotides as indicated above. We used the Clusterdraw tool to identify putative rhomboid regulatory sequences in non-*D. melanogaster* genomes (*38*). See supplementary table T1 for details of *rhomboid* and other genes' enhancer sequences and nomenclature.

All constructs were integrated into the same site on chromosome 2 (chromosomal location 51D Bloomington stock #24483). DNA microinjections were performed in-house and by Rainbow Transgenic Flies, Inc. Transgenic lines were made homozygous, and only embryos from homozygous fly lines were used for confocal microscopy.

Immunofluorescent in situ hybridization

Embryos were collected and fixed as previously described (*39*). Immunofluorescent in situ hybridization was done essentially as previously described with some modifications (*40*, *41*). All washes were done in 1 ml volume. About 50 μl of fixed embryos stored at -20^{o}C in methanol were briefly washed six times with 100% ethanol, followed by a wash in xylenes for 30 min, and lastly, six times again with 100% ethanol. The embryos were then washed four times with 50% methanol-50% phosphate buffer 0.1%-Tween 80 (PBT; 137 mM NaCl, 4.3 mM Na_2HPO_4, and 1.4 mM NaH_2PO_4) and then with PBT four times, each for 2 min with continuous rocking. Embryos were washed in (1:1, v/v ratio)

PBT/ hybridization solution (hybridization solution: 50% formamide, 5X SSC [0.75M NaCl and 75 mM Na-citrate], 100 μg/mL sonicated salmon sperm DNA, 50 μg/mL heparin, and 0.1% Tween 80) for 10 min, and then briefly in hybridization solution for 2 min. New hybridization solution was added, and the tubes were placed for 1 h in a water bath at 55°C. Previously titrated antisense RNA probes of digU-labeled *lac*Z and biotin-labeled *eve* and *sna* were heated in 65 μL hybridization solution at 80°C for 3 min and directly placed on ice for 1 min; hybridization solution was completely removed from the embryos, and the probes were added to the embryos in a final volume of 65 μL in each tube, and incubated at 55°C overnight. After incubation, 1 mL of 55°C hybridization solution was added to each tube; all tubes were rocked at room temperature for 1 min, hybridization solution was changed, and tubes were incubated for another 1 h at 55°C, followed by four washes with hybridization solution for 15 min each at 55°C and with hybridization solution and PBT (1:1, v/v ratio) two times at room temperature for 15 min. Five more washes were done with PBT for 10 min with rocking at room temperature. The embryos were washed with a blocking solution consisting of a mixture of PBT and 10% casein in maleic acid buffer (Western Blocking Reagent; Roche 11921673001) (4:1, v/v ratio). 0.5 ml of a 4:1 v/v mixture of PBT and 10X blocking solution containing primary antibodies (3 μl of sheep anti-digoxigenin, (Roche 11333089001); 1 μl of mouse anti-biotin (Invitrogen 03-3700) was added, and the tubes were rocked at 4°C overnight. Embryos were washed four times each with PBT for 15 min at room temperature. 0.4 ml of mixture of PBT and 10% casein blocking reagent and PBT (4:1 v/v), containing 8.0 μl of each secondary antibody (donkey anti-sheep Alexa 555 (Invitrogen A-21436) for detection of *lac*Z mRNA and donkey anti-mouse Alexa 488 (Invitrogen A-21202) for

detection of *eve* and *sna* mRNA) Secondary antibodies had been pre-absorbed for at least 2 hours against fixed *yw* embryos in PBT. 0.4 μl of TOPRO-3 DNA dye (Invitrogen, T3605) were also added to each vial. Tubes were covered with aluminum foil to protect them from light and incubated overnight with rocking at 4°C. Embryos were then washed with PBT four times at room temperature for 5 min. with rocking, and washed in glycerol-PBT (7:3, v/v ratio) for 2 hours until the embryos settled to the bottom of the tubes. The embryos were then resuspended in 0.4 mL glycerol-PBT (9:1, v/v ratio) and 0.2 mL of Permafluor™ mounting medium (Fisher TA-030FM), mounted on labeled slides, and covered with large rectangular Corning cover slips (No. 1.5; 24 X 50 mm). The slides were protected from light and stored flat at room temperature until the embryos were imaged.

Confocal Laser Scanning Microscopy

The Olympus Spectral FluoView FV1000 Confocal Laser Scanning Microscope (Olympus, Center Valley, PA) configured on an IX81 inverted microscope was used for capturing the confocal fluorescent images. For each scan of mounted embryos on a particular day of imaging, the same microscope settings for wild-type *rho* transgenic embryos were employed to all images to allow for direct comparison of intensities. The 488nm argon laser was used for excitation of the Alexa 488; the 559nm solid-state laser was used for excitation of the Alexa 555, and the 635nm solid-state laser was used for excitation of the TOPRO-3. Emitted fluorescence was detected using a 500–545nm band pass filter for detection of the Alexa 488, a 570–625nm band pass filter for detection of Alexa 555, and a 655-755nm band pass filter for detection of TOPRO-3. The pinhole aperture was set to 1.0 Airy unit. PMT voltage was set at maximum for

images obtained from embryos transgenic for the wild-type *rho*NEE enhancer, avoiding saturation of signal intensities. Other constructs were imaged subsequently on a single day without changing any of the microscope settings. Embryos were imaged at a scan speed of 12.9 s/scan and a Kalman average of 2. A total of 21-30 confocal images through the Z thickness were acquired for each embryo with a Z-step interval of 1.16 μm per step. CLSM image data was stored as three separate stacks and projections of images for each channel. The section dimensions were 333 mm in length and width, and 1.73 mm in depth. Fluorescence pixels were recorded as 12-bit images and stored as TIFF files. To control for overall fading of signal post-staining, wild-type *rho* constructs were stained in parallel and used to normalize overall signal intensity for each imaging day. Stained embryos were imaged within a week to minimize loss of signal.

Image Processing

All confocal microscopy images were processed in a six-step procedure involving binary image generation, rotation, resizing, background subtraction, normalization and intensity-value extraction. Binary image generation, rotation and resizing were done as described before (*42*). The area of interest for all embryos comprised a region spanning from 40%-60% egg length on the anterior-posterior axis. Ten samples, uniformly spaced, were taken from this region and plotted together from dorsal to ventral. For background subtraction, analysis of background signals from non-transgenic, *yw* flies showed a parabolic-shape (*43*), therefore a quadratic function was fit to the region of the signal representing the dorsal ectoderm, where *rho* is not expressed, and subtracted from the raw fluorescent signal. To normalize signals, values from each image were

normalized to the average peak (>95%) wild-type signal obtained during the same imaging session. This procedure allows for images to be compared for a single construct imaged on multiple days, as well as to compare intensity from one construct to another. The average intensity profiles, along with confidence intervals are given in supplementary figure S3.1.

Confocal Image Dataset

For the 59 constructs analyzed, a total of 935 embryo images were taken, with a minimum of 10 images per construct. Late stage 5 (pre-gastrulation) cellularizing embryos were used for analysis, and *eve* expression was used to select the embryos in a narrow age range. Embryos were also selected based on their rotation, so that the *rhomboid* lateral stripe was near the center of the image, with sufficient number of pixels in the dorsal region of the embryo for background estimation.

Sequence Analysis

Because there are slight differences in the reported PWMs for Dorsal, Twist, and Snail, we considered information from a variety of sources. For Dorsal, PWMs were obtained from two sources: a PWM generated by MEME analysis (with default settings) of footprinted binding sites found in FlyReg (*44*), and bacterial one-hybrid experiments (*45*). The two matrices were then averaged, and the log values calculated from this averaged matrix were used to yield a third PWM for Dorsal. For Twist, PWMs were used from two different SELEX experiments (*33*, *46*). Subsequently, a third PWM was then derived by averaging the two PWMs as described above for Dorsal. For Snail, three different sources were used: SELEX data from BDTNP (http://bdtnp.lbl.gov), SELEX

data from a previously published study (*33*), and a PWM generated by MEME analysis of footprinted binding sites found in FlyReg database (*44*)

For analysis of enhancer sequences, we used the MAST program from the MEME software suite to identify putative binding sites (*47*). The thresholds used in thermodynamic modeling were evaluated by recovery of known footprinted binding sites, although for some settings not all PWMs were able to find footprinted sites.

Quantitative data for Dorsal, Twist, and Snail

Quantitative values for Dorsal, Twist and Snail were obtained for early *Drosophila* embryo (stage 5) from a previously published study (*33*). The published data consisted of 1000 average concentrations for each protein uniformly distributed along the DV axis. Since we were only concerned with the portion of the embryo in the Ventral region, we took the region from 0 – 40 % of the DV axis and chose a subset of the 1000 data point (17 uniformly distributed data points corresponding to this region; hence a data point every 2.5% egglength) as our Dorsal, Twist, and Snail concentration gradients. The data used for modeling is given below:

Dorsal: 0.85326 0.77516 0.68914 0.59981 0.51152 0.42792 0.35175 0.28472 0.22757 0.18021 0.14193 0.11165 0.08811 0.07001 0.05618 0.0456 0.03746

Twist: 0.93224 0.88219 0.81279 0.70658 0.54216 0.34085 0.17674 0.08318 0.03873 0.01842 0.00892 0.00433 0.00208 0.00097 0.00044 0.00019 0.00008

Snail: 0.985 0.976 0.967 0.957 0.902 0.441 0.043 0.005 0.001 0 0 0 0 0 0 0 0

Structure of Models

To test different hypotheses about biochemical mechanisms of transcription factor activity on enhancers, several different schemes involving transcription factor

cooperativity and short-range repression were implemented in our modeling effort. To create models that considered the diverse cooperativity and repression ("quenching") relationships we propose, all possible pair-wise combinations of the cooperativity (15) and quenching (8) approaches were considered, generating 120 different models.

For short-range repression, we used three continuous functions (Linear-Q2, Logistic-Q3 and Exponential Decay-Q4) to describe change in repressor activity as a function of increasing distance to the nearest activator binding site.

1) Linear –

$$f\,(d)\;=\;a\;+\;bd,$$

2) Logistic Decay –

$$f\,(d)=\;ae^{-d^2/b}$$

3) Exponential Decay –

$$f\,(d)=\;a\left(\frac{1}{e^{\frac{d}{b}}}\right)$$

When implemented, a = 1 and 'b' is a model parameter for quenching functions, and 'a' and 'b' are both model parameters for cooperativity functions.

An alternative approach involved "binning" distances between activators and repressors. We fit quenching parameters (Q) for each of the bins. We also used the non-monotonic "quenching" function (Q1) derived from our analysis of short-range repression by the Giant protein in synthetic enhancer constructs (*20*).

The binned quenching schemes are described as follows. The distances between binding sites were calculated from the center of the binding sites. Because of minimal center-to-center distances between Snail and Twist or Dorsal, the actual minimal distance possible would be 11 bp in wild-type *rho* enhancer sequence.

Scheme Q5: Q1: 1-25 bp**,** Q2: 26-50 bp, Q3: 51-75 bp, Q4: 76-100 bp

Scheme Q6: Q1: 1-35 bp, Q2: 36-70 bp, Q3: 71-105, Q4: 106-140 bp

Scheme Q7: Q1: 1-45 bp, Q2: 46-90 bp, Q3: 91-135, Q4: 136-180 bp

Scheme Q8: Q1: 1-10 bp, Q2: 11-20 bp… Q9: 81-90 bp, Q10: 91-100 bp

We considered two different ways of estimating cooperativity between transcription factors: heterotypic (between Dorsal and Twist) and homotypic (Dorsal-Dorsal, Twist-Twist, or Snail-Snail). We tested three different continuous functions (Linear-C1, Logistic-C2 and Exponential Decay-C3), which were parameterized with a single parameter for all homotypic interactions, and a separate value for Dorsal-Twist cooperativity. Additional models with “binned” distances were also considered. For each of the binned schemes, we used a simpler form in which all homotypic interactions are parameterized with the same value, and a more complex form where each type of protein interaction for a given bin size receives a distinct parameter. Each of these schemes therefore generates two model forms – binned and protein-binned respectively.

Schemes C4 and C10: C1: 1-25 bp, C2: >25 bp

Schemes C5 and C11: C1: 1-50 bp, C2: >50 bp

Schemes C6 and C12: C1: 1-75 bp, C2: >75 bp

Schemes C7 and C13: C1: 1-50 bp, C2: 51-100 bp, C3: >101 bp

Schemes C8 and C14: C1: 1-60 bp, C2: 61-120 bp, C3: >121 bp

Schemes C9 and C15: C1: 1-70 bp, C2: 71-140 bp, C3: >141 bp

Parameter Estimation

A global parameter estimation strategy, CMA-ES (Covariance Matrix Adaptation - Evolutionary Strategy) was applied to estimate the parameters (*48*, *49*). Root mean square error (RMSE) was used as a measure of performance of different cooperativity and quenching schemes, as described before (*20*). Due to the stochastic nature of starting points and fixed maximum number of runs for CMA-ES, estimations were run five times, which was empirically found to be sufficient to produce similar, minimal RMSE values for at least three of the runs in over 56% of cases.

Cross Validation

Systematic cross-validation

Constructs were divided into five sets based on the type of mutation as follows:

Dorsal site knockouts: Constructs 2,3,6,7,8,11,12,15,16

Twist site knockouts: Constructs 4,5,17

Dorsal and Twist site knockouts: Constructs 9,10,13,14,18,19,20,21

Snail site knockouts: Constructs 22-33

bHLH site knockouts: Constructs 34-38 (see supp. Table T1 for construct details).

Parameter estimation was performed using selected 24 models while leaving out data from each of the five sets of constructs. Expression was subsequently predicted for all 38 constructs using parameters obtained, and RMSE over the constructs left out as well as over all 38 constructs was used to analyze the effects of data provided by each set of constructs to the model.

Random, Five-fold cross-validation

The 38 constructs to be fitted were separated into 5 randomized partitions of size eight (three partitions) and seven (two partitions). The partitions were computer-generated using the Python random.shuffle() method, which is based on the Mersenne Twister algorithm (*50*). This process was repeated five times to give five different partitioning schemes. All 38 constructs were then predicted using parameters from each run, and average RMSE of the constructs left out was considered.

Sensitivity Analysis

Sensitivity analysis was performed only for 24 selected models as previously described (*51*). Uninformative parameters, i.e., those with empty bins were excluded from the analysis.

Scoring of predictions from enhancers (cloned into pHonda1)

For neurectodermal enhancers, a four-point scheme was applied to score Snail repression as well as neurectodermal activation. Snail repression was measured at nucleus 4. Snail repression was scored as:

1. wild-type repression (expression below 0.1)
2. some loss of repression (expression 0.1-0.3)
3. higher loss of repression (expression 0.3-0.5)
4. very poor repression (expression 0.5-1.0)

Neurectodermal activation was scored at the peak in a four-point scheme:

1. Difference between predicted and measured peak expression is less than 0.2
2. Difference between predicted and measured peak expression is between 0.2 and 0.5

3. Difference between predicted and measured peak expression is between 0.5 and 0.7
4. Difference between predicted and measured peak expression is greater than 0.7

For mesodermal enhancers, scoring was done on a five-point scale for activation and four-point scale for Snail activity. The activation score is given below:

1. Mostly mesoderm activation, difference between predicted and measured peak expression is less than 0.2
2. Mostly mesoderm activation, difference between predicted and measured peak expression is between 0.2 and 0.4
3. Mostly mesoderm activation, difference between predicted and measured peak expression is between 0.5 and 0.7
4. Low mesoderm activation, high neurectoderm activation, low dorsal ectoderm activation
5. Low mesoderm activation, high neurectoderm activation, high dorsal ectoderm activation

Snail repression scale is given below:

1. No Snail activity, putative mesoderm activation is equal to 1.5 times peak neurectoderm expression
2. Some Snail activity, mesoderm activation = peak neurectoderm expression
3. High snail activity, mesoderm activation< neurectoderm activation
4. Highest snail activity, low expression in mesoderm (<0.1 intensity value)

Prediction of Dorsal regulon enhancers

Selected peaks for Dorsal, Twist and Snail binding were obtained from a previously published study (*30*). Sequence data for these peaks was obtained from UCSC Genome Browser (genome.ucsc.edu) using the April 2004 (BDGP R4/dm2) assembly. Binding sites were predicted using following PWMs: Dorsal (Flyreg), Twist (Average), Snail (Flyreg). Expression patterns for genes associated with peaks were obtained from BDGP (*52*) and FlyExpress (*53*), and were categorized as expressing in mesoderm (M), neurectoderm (N), mesectoderm (ME) and mesoderm and neurectoderm (MN). Since our dataset and predictions are categorized into 18 nuclei on dorsal-ventral axis, we established following domains of different tissues – Nuclei 1-6 as mesoderm, nuclei 7-15 as neurectoderm, nuclei 7-9 as mesectoderm and nuclei 1-15 as mesoderm and neurectoderm. All expression patterns were thresholded at 0.1 intensity value. Predictions were categorized into three categories – Good – expression in all the correct nuclei and misexpression in less than 6 nuclei, average – expression in all the correct nuclei and misexpression in less than 7 nuclei, bad – expression not in categories of good or bad.

Results

Training of thermodynamic models on perturbation datasets for genome-wide prediction of regulatory information

Our experimental strategy was to first obtain a high quality dataset on a regulatory element that interacts with well-characterized transcriptional regulators (Fig. 3.1A-C). Knowledge about binding site preferences and expression of the proteins is critical. Then, the systematic perturbation and quantitation of the regulatory element provides a broad basis on which to test diverse models for protein activity on this enhancer, and

fitting of each model provides a possible tool to predict the activity of diverse members of the regulon (Fig.3.1 D-G). To develop an extensive quantitative data set required for modeling Dorsal transcription factor action, we focused on the neurectodermal enhancer (NEE) of the *rhomboid* (*rho*) gene, which encodes a serine protease important for EGF signaling in *Drosophila* development. This gene is first expressed as two lateral stripes in the presumptive neurogenic ectoderm of fly embryo (*54*, *55*). A minimal 318-bp enhancer located 1.7 kb 5' of *rho* drives expression in presumptive neuroectoderm, under cooperative action of the Dorsal and Twist activators (Fig. 3.1A). Expression is excluded from the mesoderm (ventral region) by the Snail repressor (*35*). These transcriptional regulators of the *rho* gene also have an extensive role in dorsal-ventral patterning of the early embryo. From gene expression and in vivo binding analysis, it is estimated that Dorsal, Twist and Snail coordinate the expression of some 100 genes at this point in development (*30*, *56*).

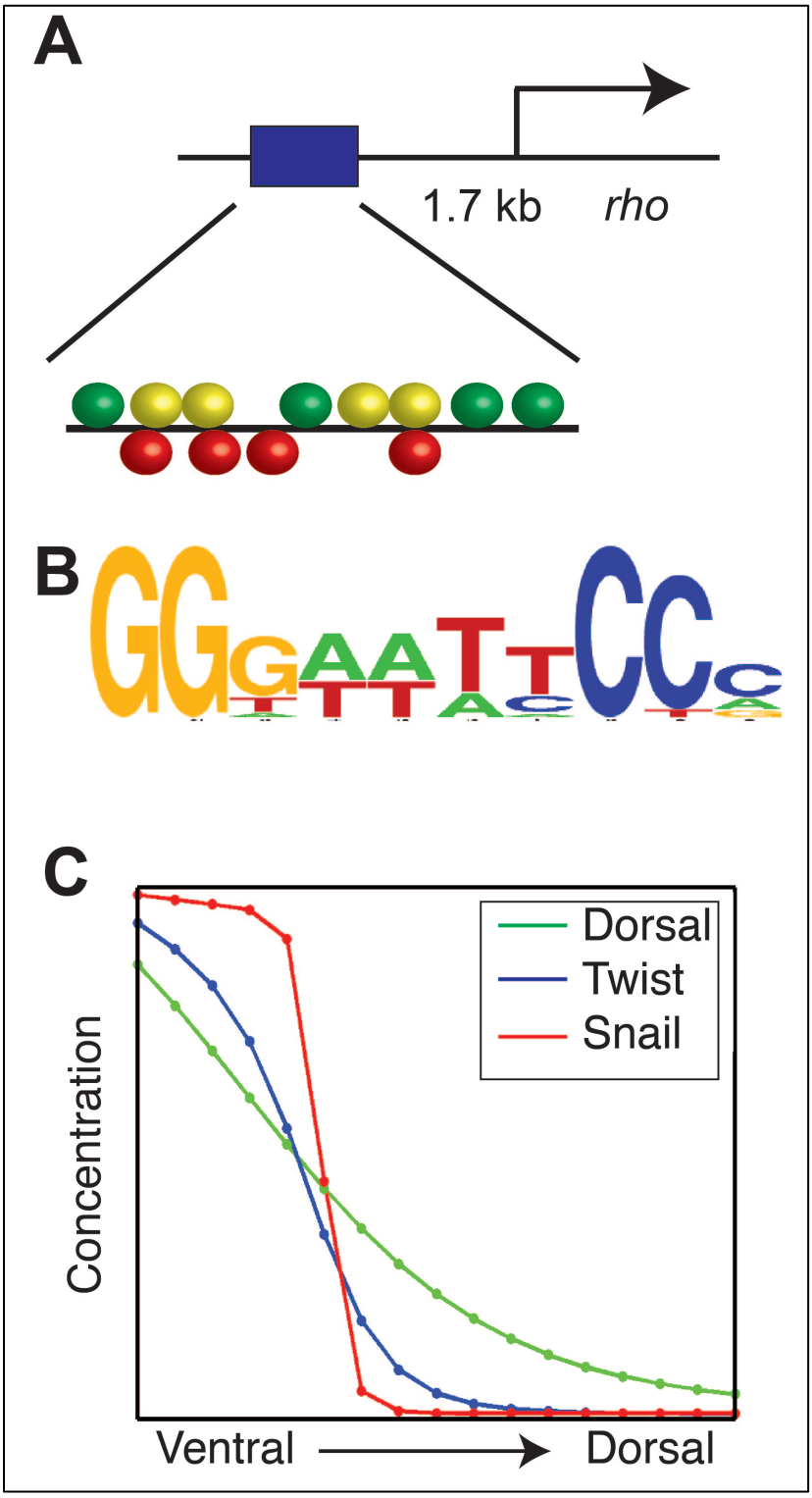


Figure 3.1. Schematic overview of study. Panels A-D describe the input data for study, panel E describes the acquisition and processing of quantitative data, and panels F-G describe thermodynamic modeling, parameter estimation and prediction of gene

Figure 3.1 cont'd

expression from Dorsal regulon enhancers (A) The minimal *rhomboid* (*rho*) enhancer is 318 bp in length and contains binding sites for Dorsal, Twist, Snail, and putative bHLH proteins. The enhancer is located 1.7 kb upstream of *rho* gene. (B) Motif information for above factors was obtained from various sources and used for scoring binding sites. (C) Quantitative information on levels of Dorsal, Twist and Snail proteins in the early fly embryo is used for modeling. (D) Eight different categories of mutations were made on *rho* enhancer, viz., (i) single activator site, (ii) two Dorsal sites, (iii) one Dorsal and one Twist site, (iv) two Twist sites, (v) all repressor sites, (vi) three repressor sites, (vii) two repressor sites, and (viii) bHLH-factor binding sites combined with Dorsal or Twist sites. Mutated sites are indicated by crosses. (E) (i) Confocal Laser Scanning Microscopy (CLSM) was used to image *lac*Z mRNA expressed by *rho* enhancer (red), along with

sna (green horizontal stripe) and *eve* mRNA (green vertical stripes). (ii) Data from *lac*Z mRNA was plotted from ten equally-spaced intervals from 40-60% egg length (EL) of embryo. (iii) Background subtraction was performed based on signal intensities in dorsal regions, where activator concentrations are the lowest. Image normalization was carried out using wild-type *rho* constructs stained and imaged in parallel. This normalized, background-subtracted data was then averaged to give a single plot of expression values for an image (red line). (iv) Data from at least 10 embryos was pooled to calculate average expression levels of a construct. Dotted lines indicate standard error of mean. (F)(i) Thermodynamic models calculate gene expression based on equilibrium probabilities of binding of activator proteins on enhancer. In a hypothetical enhancer shown here with one Dorsal (D) and one Snail (S) site, probability of mRNA expression is proportional to all successful states (numerator) divided by all possible states (denominator). (ii-iii) Different functions of cooperativity and quenching, including continuous (ii) or binned (iii) were used to fit the model to the quantitative data. Dotted line indicates measured expression data, while red line indicates model fit. (G) (i) 120 different quenching and cooperativity model combinations were trained on the dataset containing 38 *rhomboid* enhancer constructs. (ii) Parameters derived from the model training were used to predict expression pattern for putative enhancer sequences enriched in binding by Dorsal, Twist and Snail identified in ChIP-chip experiments. Shown here is a schematic binding genomic binding peak (top) and expected gene expression from the DNA element (bottom).

Figure 3.1 cont'd

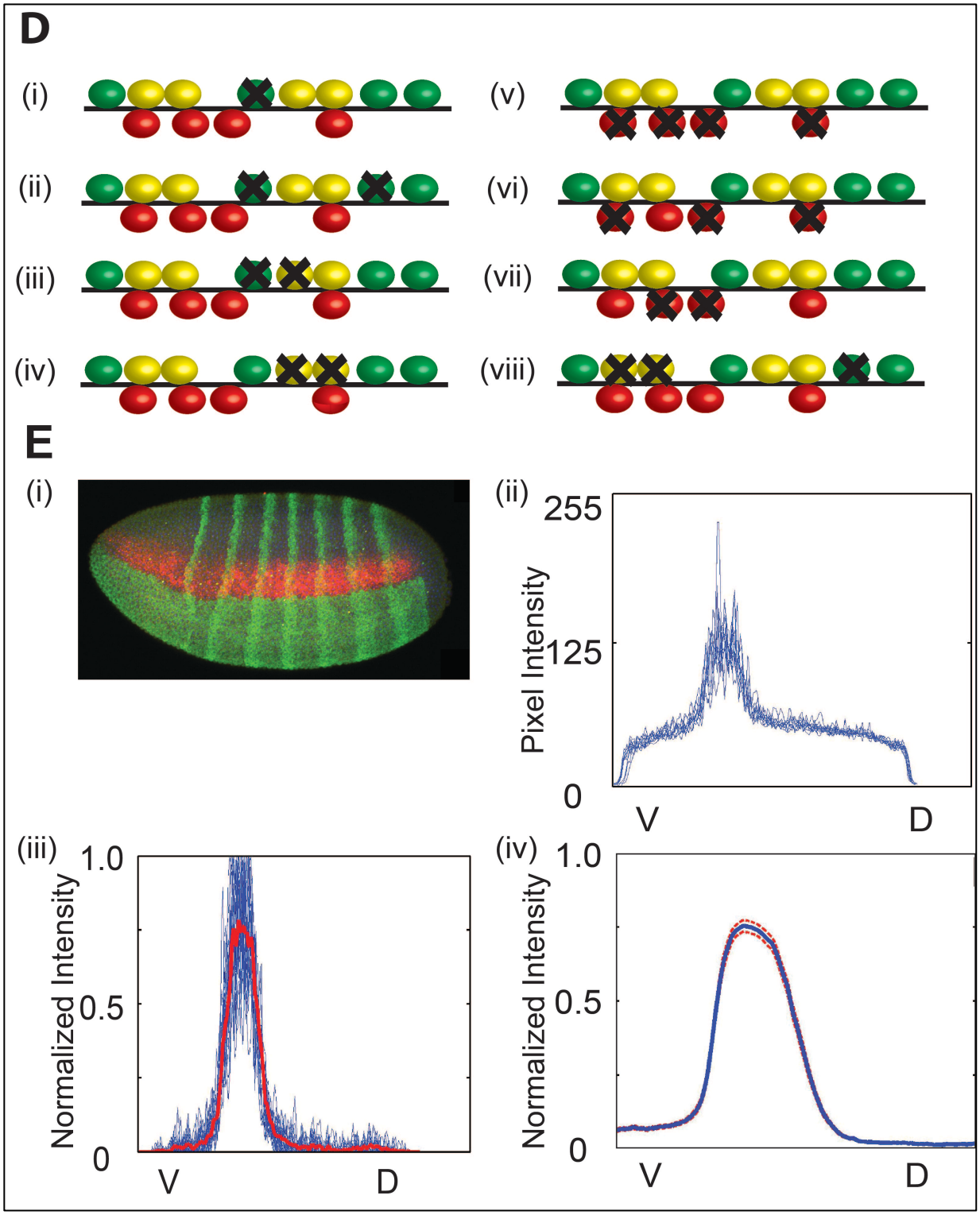
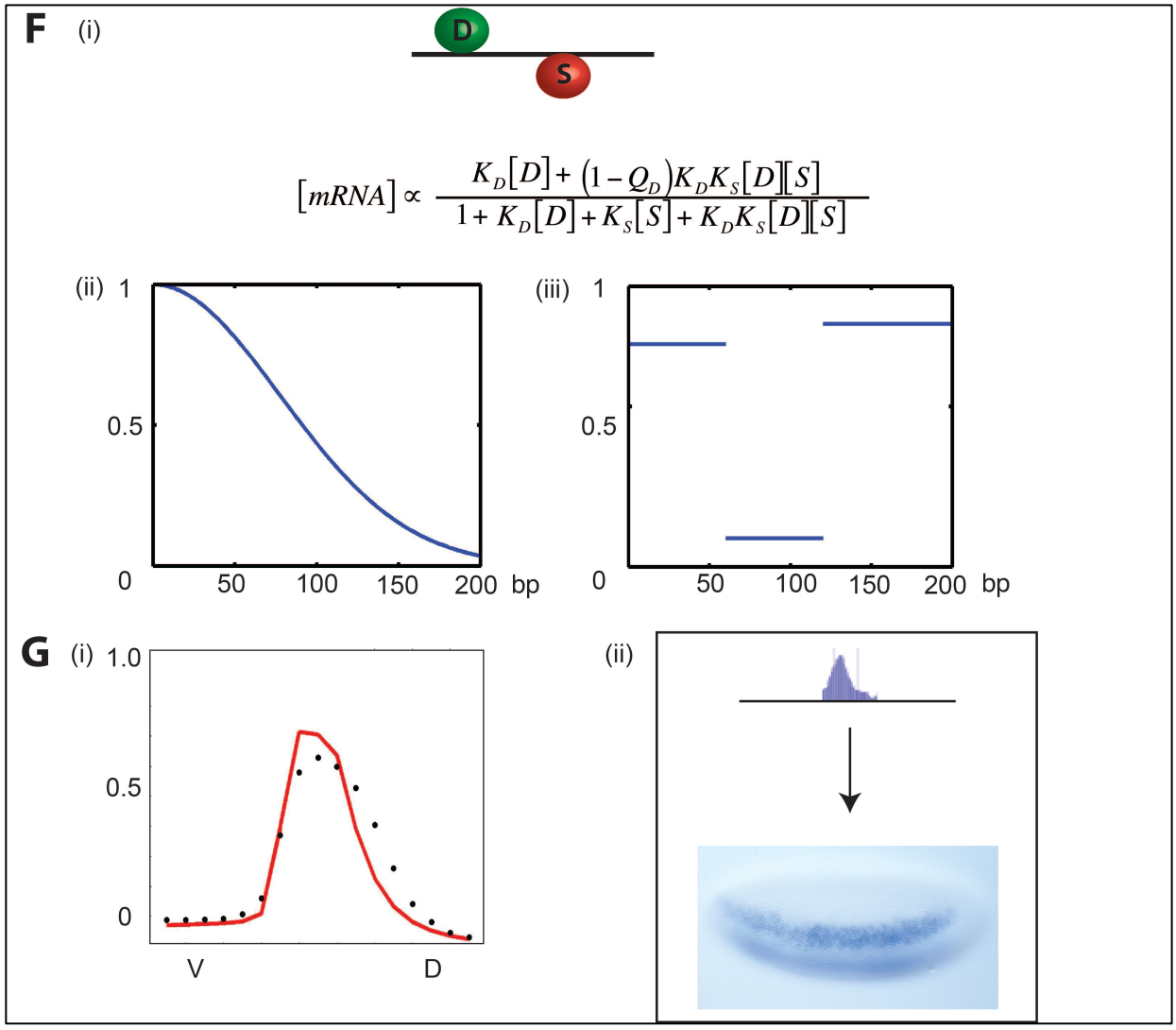


Figure 3.1 cont'd



Perturbation analysis of *rho* enhancer

There are four Dorsal, two Twist and four Snail binding sites in this enhancer identified by DNaseI footprinting. Dorsal and Twist are transcriptional activators that show cooperative interactions in vitro and in vivo (*57*, *58*). Snail acts as a short-range repressor, with the capacity to interfere with activators located within ~100 bp of a Snail binding site (*59*). We generated 37 variations of the wild-type *rho* sequence in which each Dorsal and Twist activator sites was removed by site-directed mutagenesis individually, or Dorsal and Twist sites in a pairwise manner (Fig.3.1D, Fig. 3.2 and Supplementary Table T1). Such mutagenesis had previously been demonstrated to severely impact activity of the enhancer (*35*). Single Snail sites were previously shown to be sufficient to mediate some level of mesodermal repression, therefore we made constructs in which two, three, or all four of the Snail sites were removed (*59*). All *rho* enhancers were cloned into the pHonda1 vector, and were integrated into the same position on the second chromosome using site-specific recombination (*36*). We measured the transcriptional output from the enhancers using Fluorescent in Situ Hybridization (FISH) followed by confocal laser scanning microscopy (Fig.3.1E). This quantitative assay is sensitive to differences in transcript levels (*20*, *21*, *33*, *42*). We then used an image-processing pipeline to extract quantitative gene expression data from embryos of the same late-blastoderm stage. Quantitative data from a minimum of ten embryos was normalized and combined to provide average expression patterns for each variant of the *rho* enhancer. In total, 935 images were analyzed (Fig. 3.2, Supplementary Fig. S3.1).

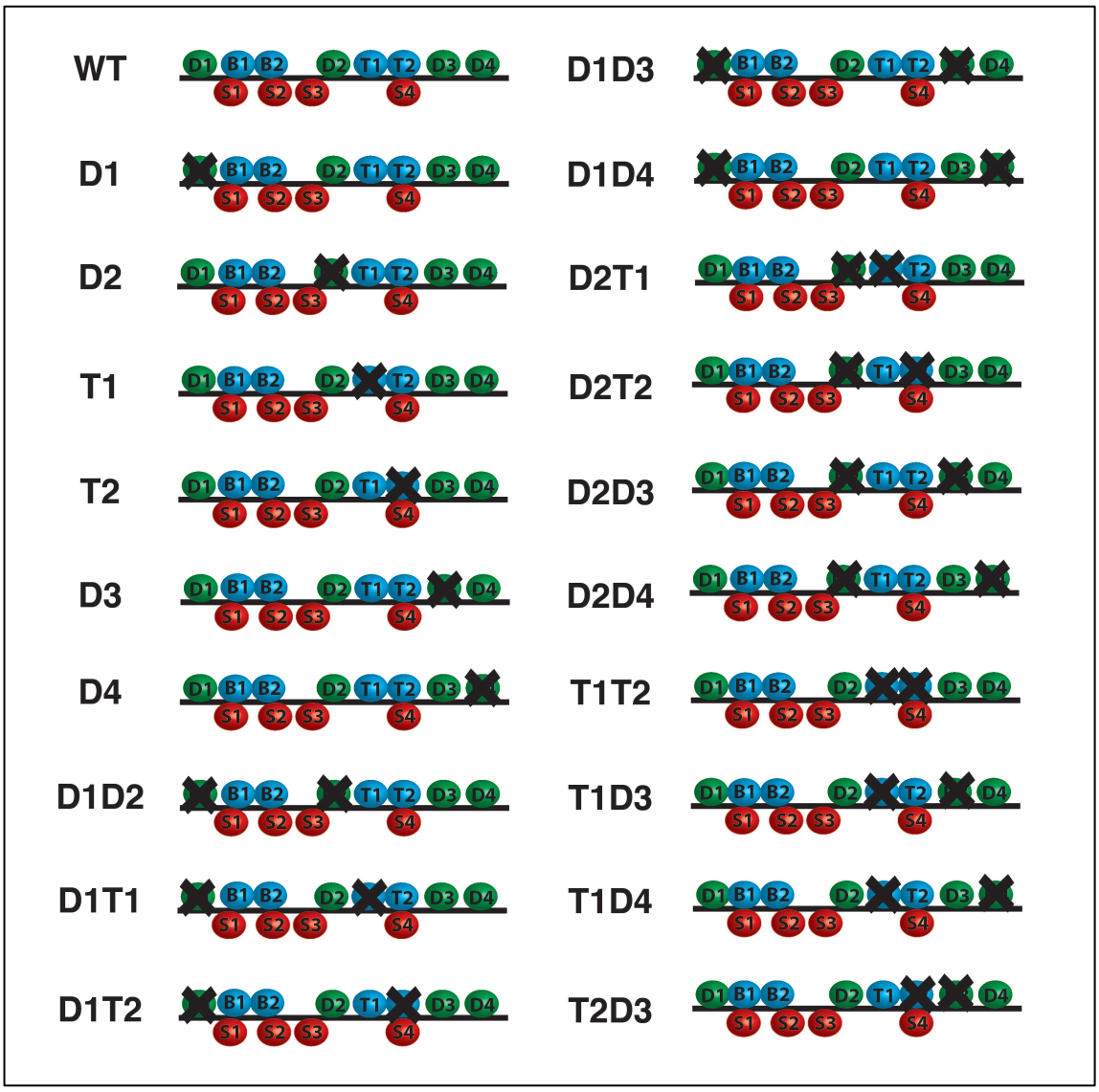
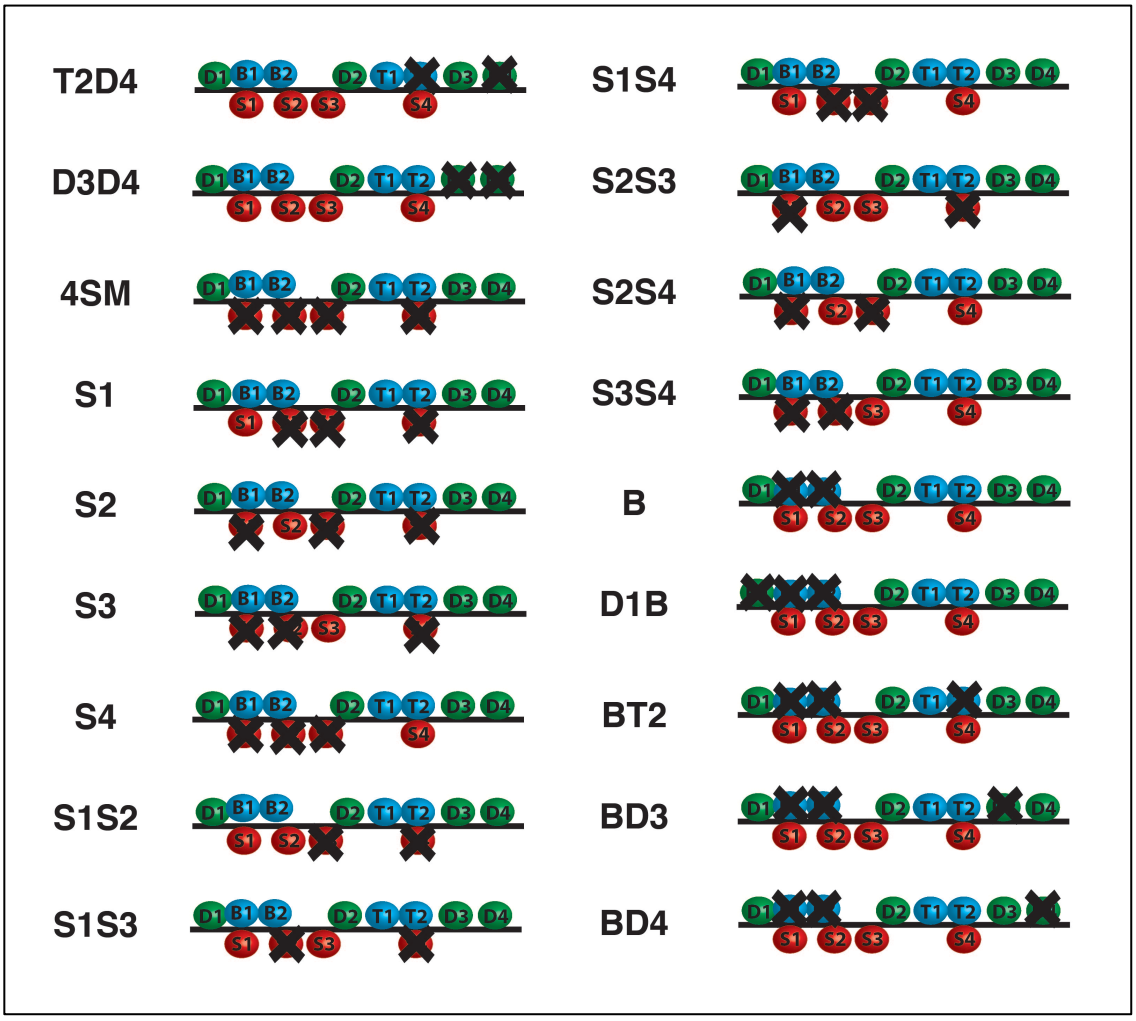


Figure 3.2. A schematic illustration of *rho* enhancer constructs. Green and red circles denote footprinted binding sites for Dorsal (D) and Snail (S). Blue circles denote binding sites for bHLH (B) factors and Twist (T). Binding sites mutated in each construct have been marked with an "X".

Figure 3.2 cont'd



Mutation of any single Dorsal or Twist activator binding site resulted in measurable reduction of peak intensity and retraction of the *rho* stripe from the dorsal region, where activators Dorsal and Twist are present in limiting concentrations (*33*). Strikingly, despite the differences in predicted binding affinities and relative positions of the motifs, the elimination of any site individually had similar quantitative effects, reducing gene expression to approximately 60% of the peak wild-type level (Fig.3.3A, Supplementary Fig. S3.1). Loss of Twist site 2 led to a slightly greater reduction in expression than other sites. Mutation of any combination of two Dorsal or Twist activator binding sites further reduced expression by greatly varying amounts, with some constructs showing only ~18-25% of the wild-type expression (for construct with Dorsal sites 2+3 or Twist2 and Dorsal3 mutated), while others showed only slightly decreased expression over the single site mutation, about 50% of wild-type (for Dorsal 3+4 or Dorsal1 and Twist1 mutant constructs) (Fig.3.3B-C). Overall, the double activator site mutagenesis revealed a complex picture of the contributions of activator sites to gene expression. The most severely compromised expression was observed with the D2D3 mutant (Fig. 3.2 and 3.3), but other combinations involving one of these sites were less dramatically impacted. We hypothesize that the variable effects of different pairwise mutations, as opposed to the rather similar effects of individual site loss, indicates that there are multiple biochemical events occurring on the enhancer with different limit points. Loss of binding by two proteins that are involved in a common vital function would thus be particularly devastating.

In contrast to the perturbation of Dorsal and Twist elements, removal of Snail repressor binding sites revealed stark differences in the significance of individual motifs

for overall activity. Previous studies indicated that placing one or two Snail sites placed inside or downstream of this enhancer is sufficient to mediate repression of the *rho* NEE, therefore we tested constructs having one or two Snail sites left intact. Mutation of all four Snail sites caused pervasive expression in the mesoderm, as expected, while constructs with a single intact Snail 2 or 3 motif showed substantial but not complete repression. Snail 1 and 4 motifs were not nearly as effective at mediating repression, although these sites are as strong as Snail2 and Snail3 sites (Fig. 3.3D). Interestingly, Snail1 is proximal to just a single footprinted Dorsal site, which may explain its poor repression efficiency. Snail2, Snail3 and Snail4 are close to at least four activator sites. However, Snail4 and Twist2 sites overlap, thus competition for binding may reduce repression efficiency (Fig. 3.3D-E). These results are in agreement with a previous study, that showed that a single Snail2 site placed 50 bp upstream of Dorsal1 site was not as effective in mediating repression as the Snail2 site placed at its endogenous location or 50bp 3' of Dorsal4 (*59*).

We also investigated the significance of two E-box motifs located between Dorsal 1 and 2 sites. These motifs do not bind Twist in vitro, but have been reported to bind to the *lethal of scute* T3 protein in vitro (*35*). Extensive scrambling of both sites caused a slight decrease in peak expression; interestingly, this was the only enhancer variant that exhibited a broader expression pattern. Simultaneous disruption of the Twist 2, Dorsal 3 or Dorsal 4 sites further reduced peak gene expression, with concomitant narrowing of the stripe. We hypothesize that the mutated bHLH sequences are less able to respond to a factor distributed on the ventral-dorsal gradient such as T3, but are better able to interact with a broadly distributed activator, such as the Daughterless bHLH protein.

Thermodynamic modeling of *rho* enhancer dataset

The generation of a high quality perturbation data set represents the first step in discerning possible enhancer grammar rules that describe the activity of Dorsal, Twist, and Snail in general. Direct examination of the data described above showed that the interrelationships among activator and repressor sites were complex, thus we created quantitative models to tease out possible interactions among the transcription factors binding to the *rho* enhancer. Based on prior findings pointing to the critical role played by transcription factor cooperativity, we focused on models that incorporated a variety of conceptions of activator-activator cooperativity, as well as different distance-dependent interactions between repressors and activators. We systematically tested continuous and step functions to determine possible distance relationships between factors. In all, we tested fifteen formulations for cooperativity, which involved in some cases estimating a single parameter that described general cooperative effects, and in others individual parameters for each sort of protein-protein interaction. Eight different formulations for repression ("quenching") were employed; because Snail has been demonstrated to

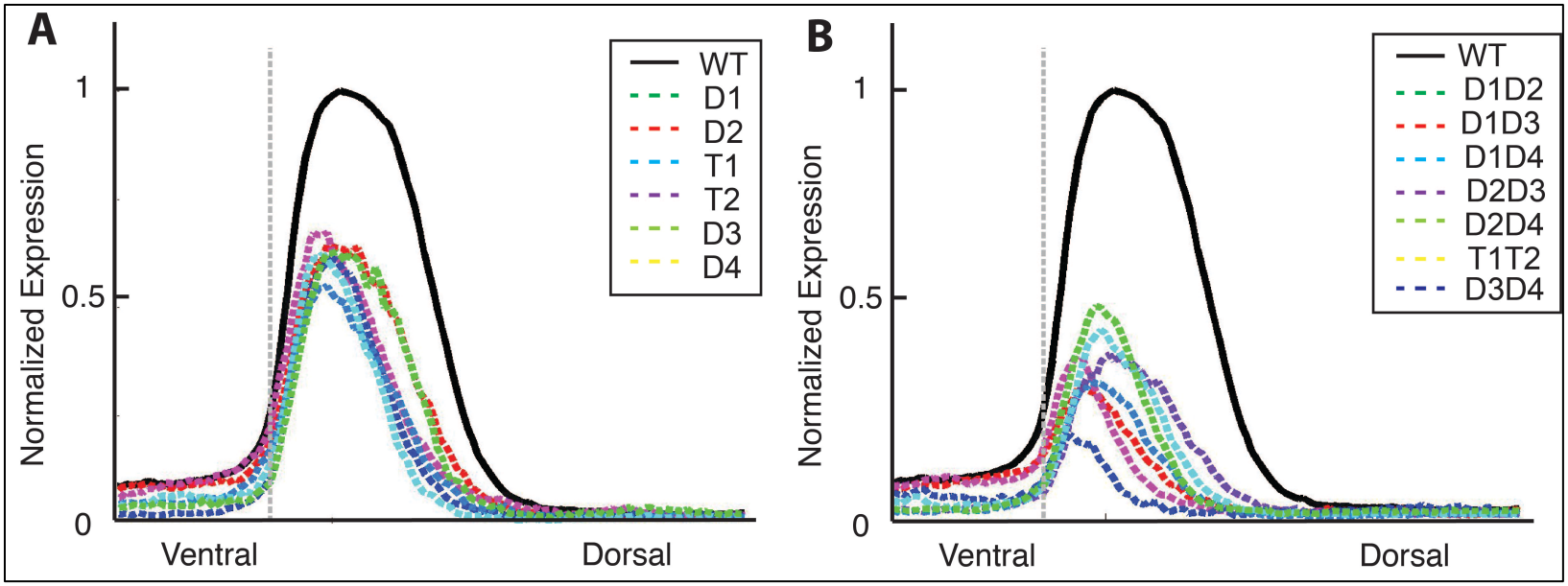


Figure 3.3 Quantitative effects of different mutations in binding sites on gene expression. Panels A-F show how mutations in different combinations of binding sites of Dorsal, Twist, Snail or bHLH-factor binding sites affect gene expression (see supplementary table T1 for construct nomenclature). (A) Effect of mutation of single activator site (Dorsal or Twist) on gene expression, shown here as a plot of *lac*Z expression from ventral to dorsal. Expression of constructs with mutations is shown, indicated in the box. Panels B-F show expression of constructs with mutations in two homotypic activator sites (B), two heterotypic activator sites (C), three or four Snail repressor sites (D), two Snail sites (E) and bHLH-factor binding sites (F). The expression driven by the wild-type *rho* enhancer is shown as a black line, and has been normalized to give a maximum expression value of 1. The grey dotted vertical line denotes boundary of *sna* expression.

Figure 3.3 cont'd

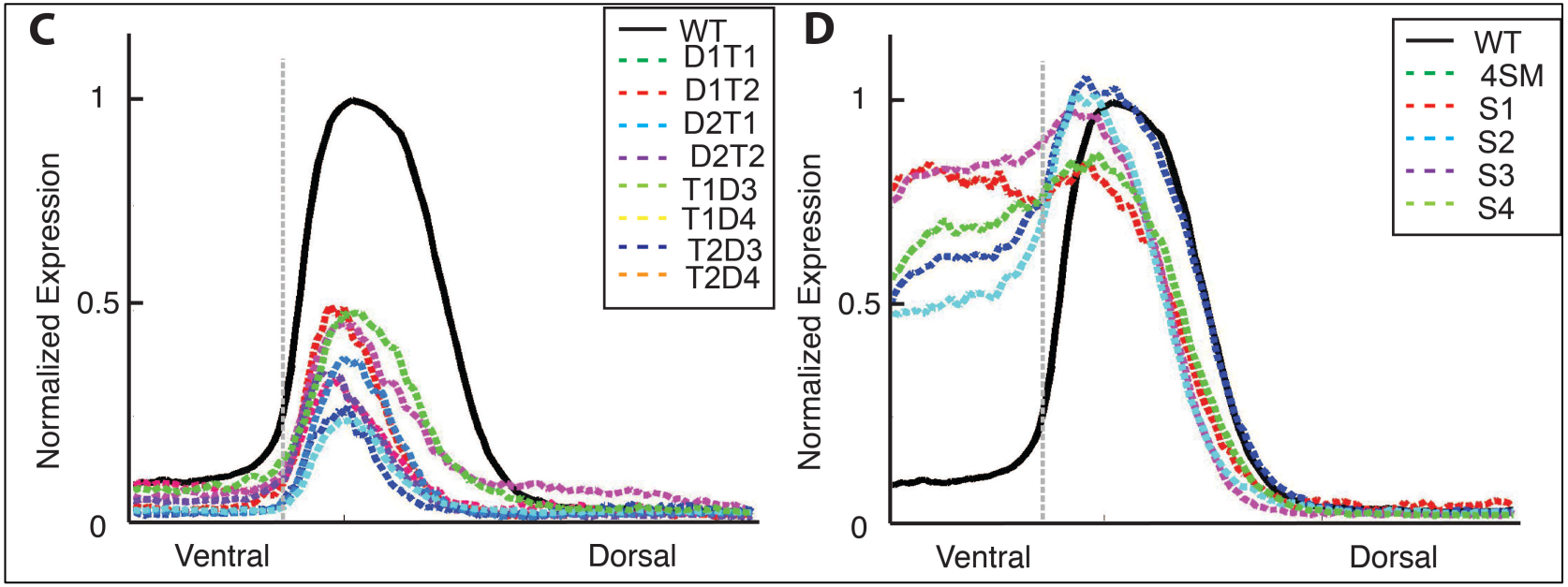


Figure 3.3 cont'd

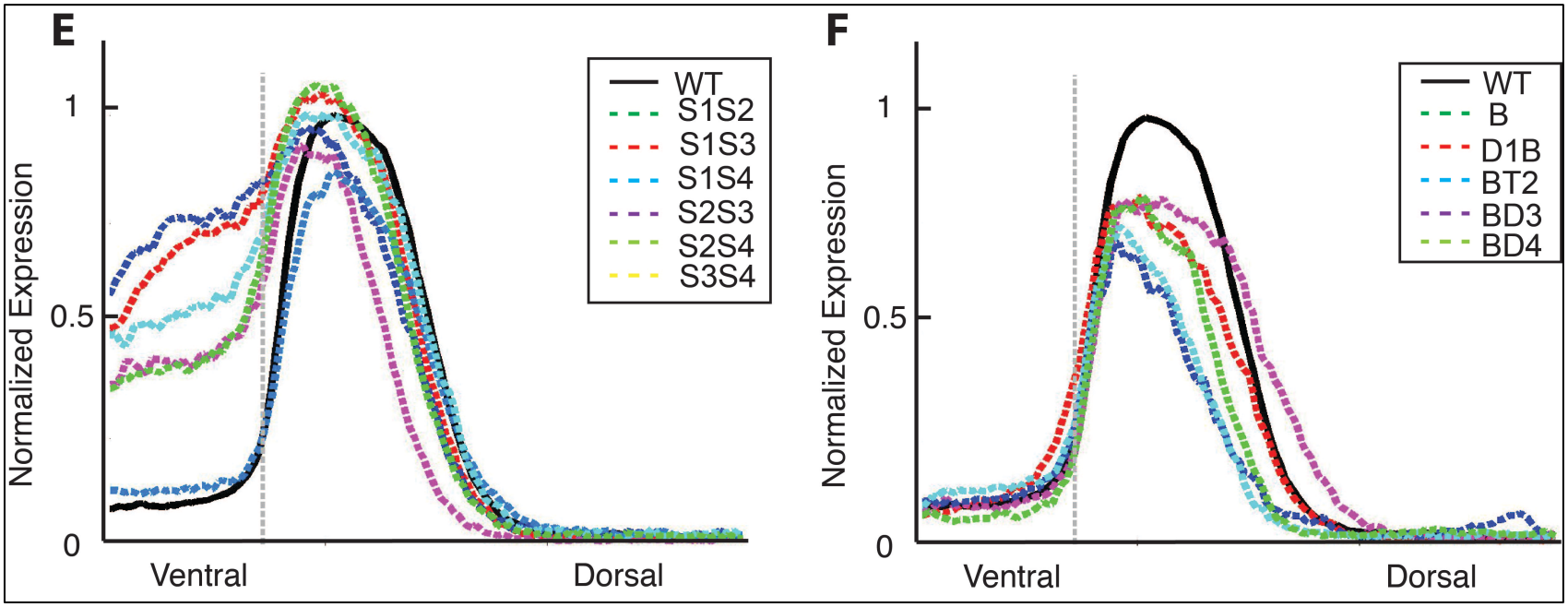


exhibit short-range repressive effects, we included descriptions that allowed for fine-grained distance differentiations. Combining these two types of formulations results in 120 different models (see Materials and Methods), the simplest of which encompassed 7 parameters, and the most complex 35 parameters.

To systematically test various hypotheses relating to interactions of activators and repressors on the *rho* enhancer, we trained each of the 120 models on the quantitative expression of the 38 variations shown in Fig. 3.2. Parameters were estimated using CMA-ES, a global genetic algorithm, and overall performance was calculated from the fit to all constructs, using Root Mean Square Error (RMSE) as the objective function. The performance of the different models as a function of average RMSE is shown in Fig. 3.4A. Global RMSE values varied from 0.071 to 0.265, which reflected in the best cases excellent agreement with measured values, and in worst cases complete inability to correctly predict Snail repression and Dorsal/Twist activation. Certain global trends are readily apparent from the inspection of the heat map displaying model performance (Fig. 3.4A). First, cooperativity models of similar structure tended to perform similarly, which is seen by examination of rows representing different cooperativity schemes. Rows 1-3 represent different continuous functions (linear, logistic, and exponential decay) fit with two parameters; overall these had higher RMSE values than Rows 4-9, which use a variety of different “bin” sizes to distinguish cooperativity relationships. These models in turn were surpassed by a set of binned models in which cooperativity is broken out so that different values are assigned for each homotypic interaction (rows 10-15).

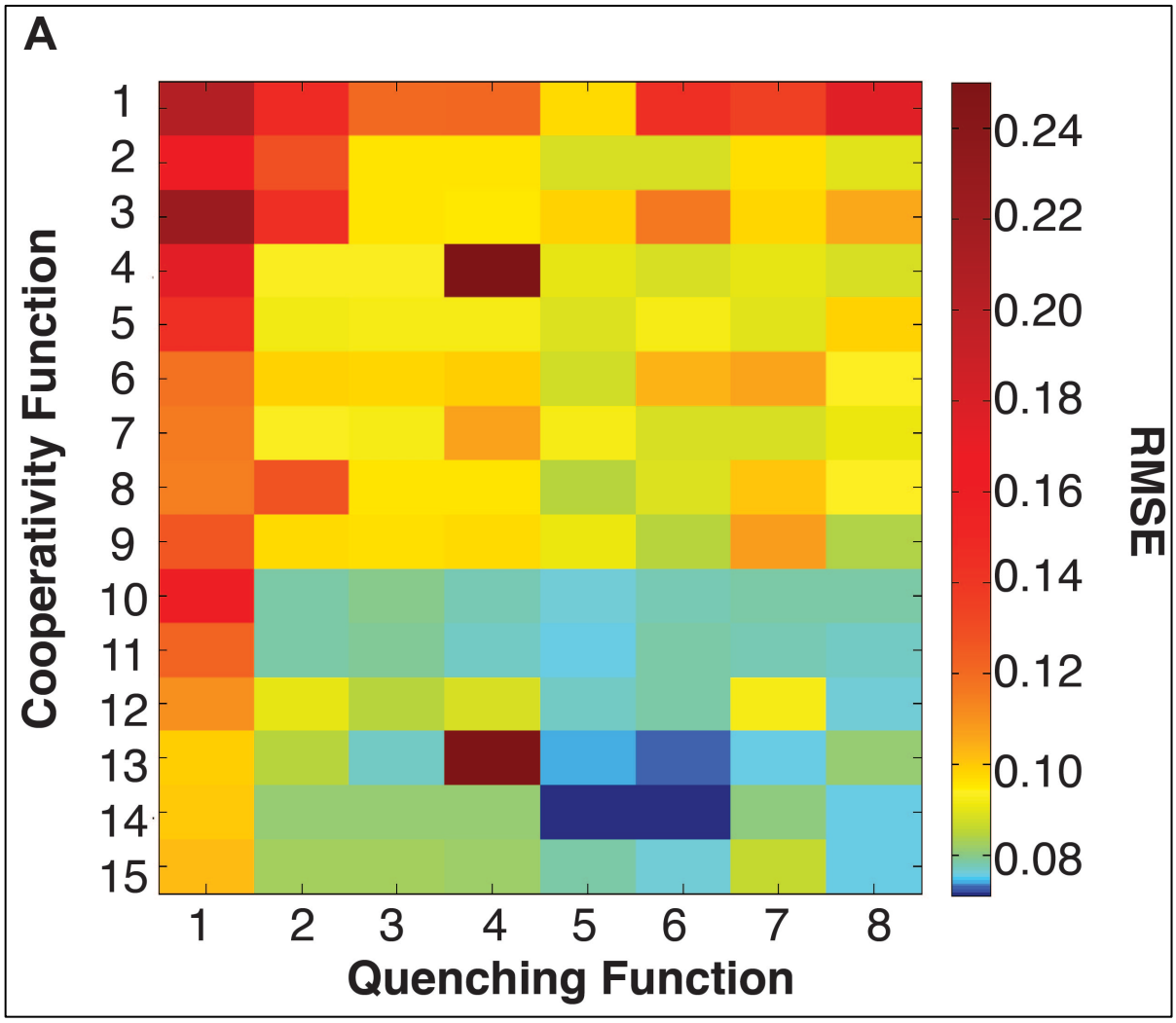


Figure 3.4. Global and construct-specific performance of 120 models on quantitative dataset. (A) The heat-map shows the performance of 120 different model combinations on the *rho* enhancer dataset. The color bar on the right indicates the RMSE score. (B) The heat-map shows the performance of 120 models on all 38 *rho* constructs. (C) Individual construct-fits for three models (C14Q5, C8Q7, and C4Q4) and five enhancer constructs (DMRW, DMRT2, DMRD1T2, DMRS2 and DMRS2S3) are shown. Vertical axis represents expression intensity and horizontal axis represents expression along the dorsal-ventral axis. Dotted lines depict experimentally measured expression and red lines indicate model fits.

Figure 3.4 cont'd

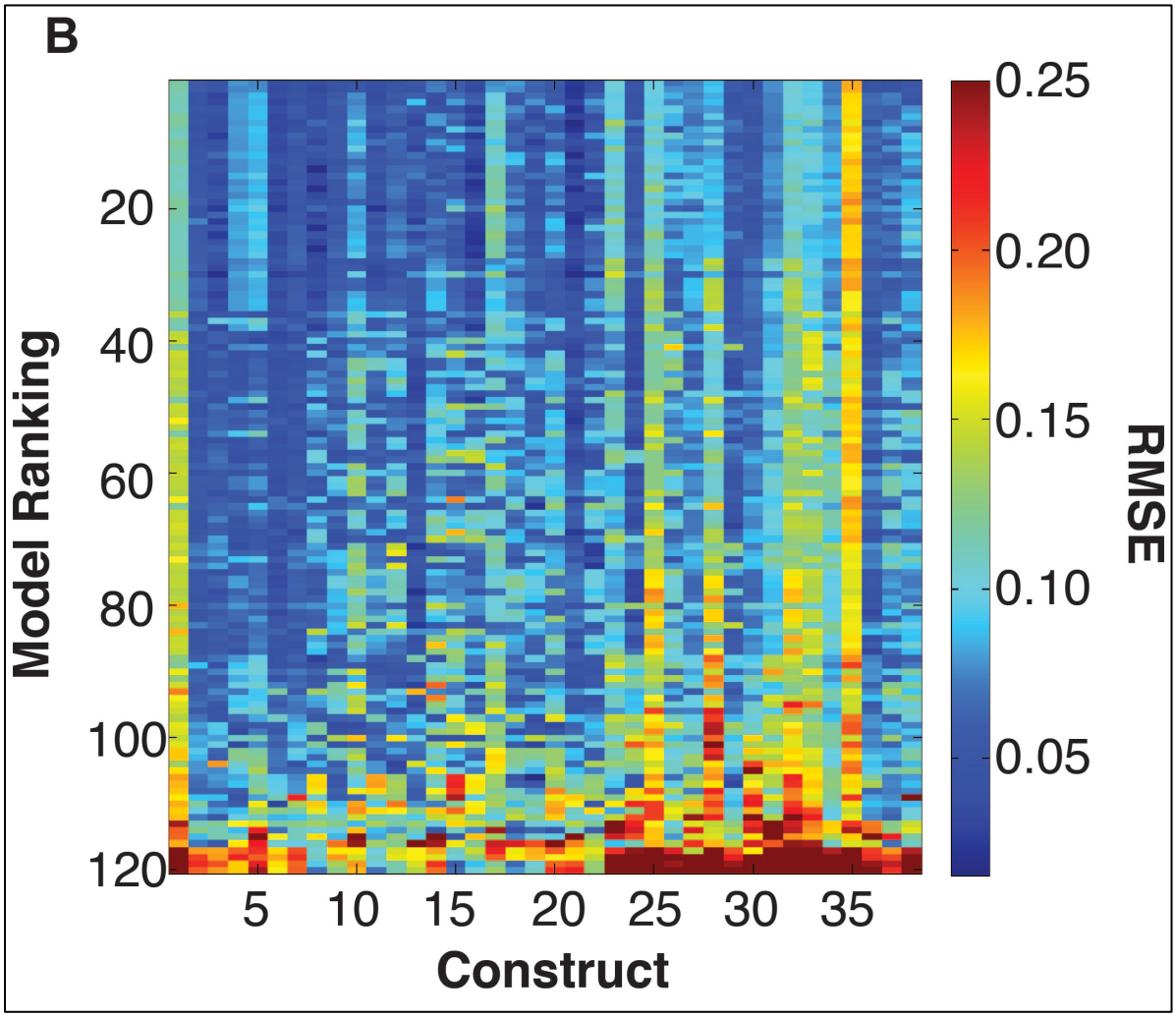


Figure 3.4 cont'd

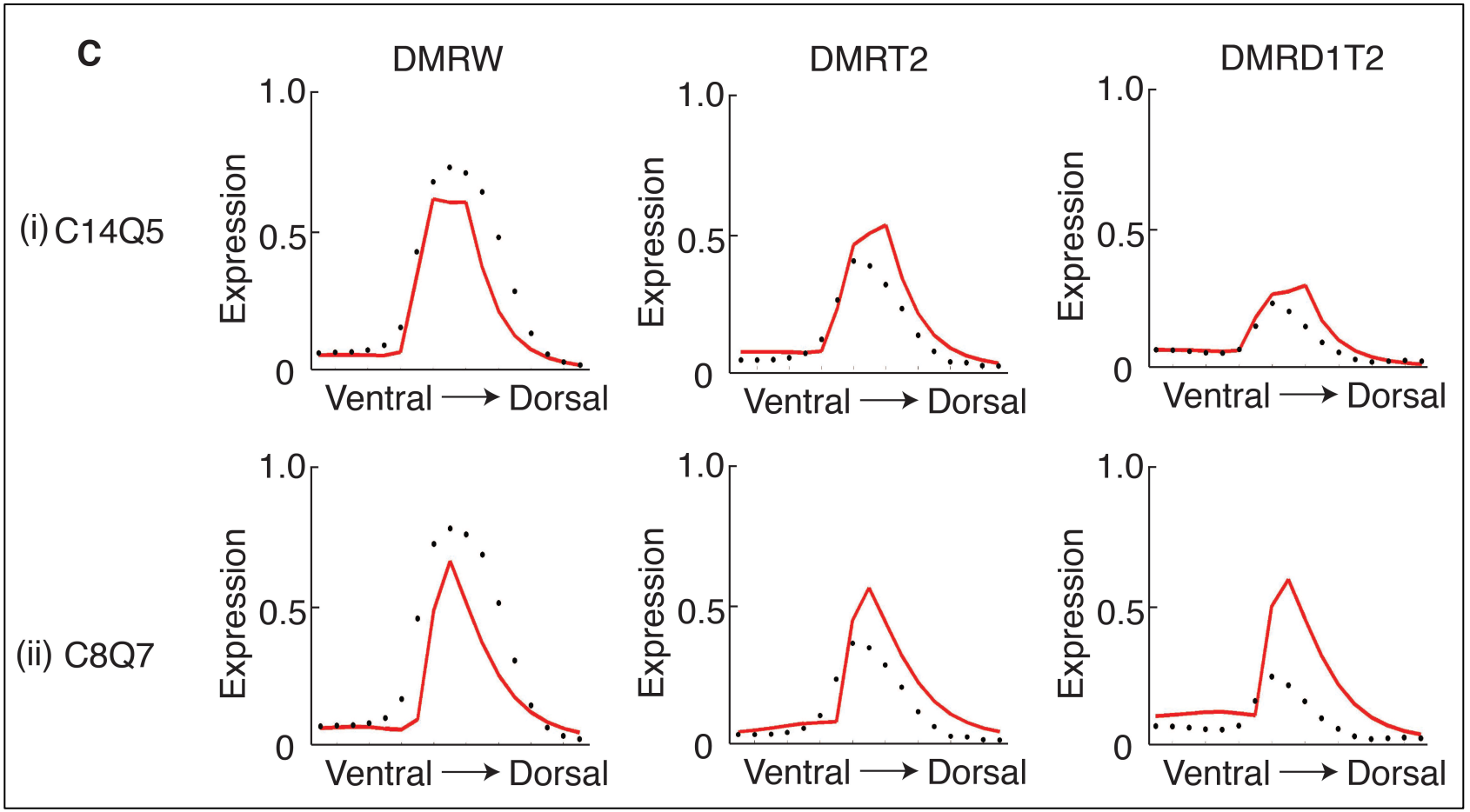


Figure 3.4 cont'd

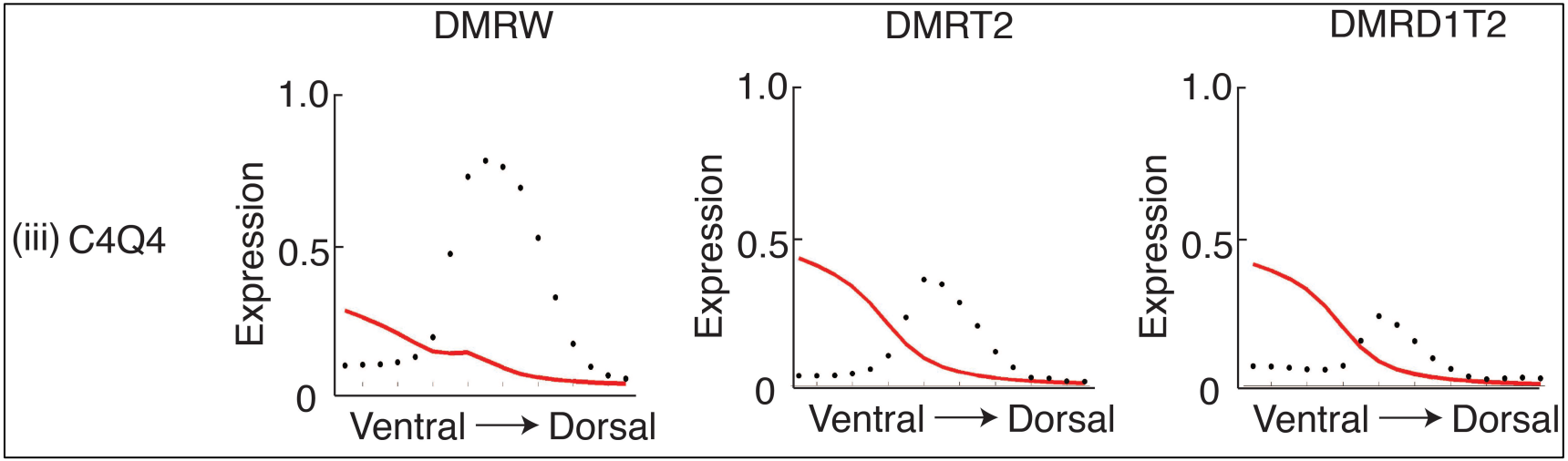
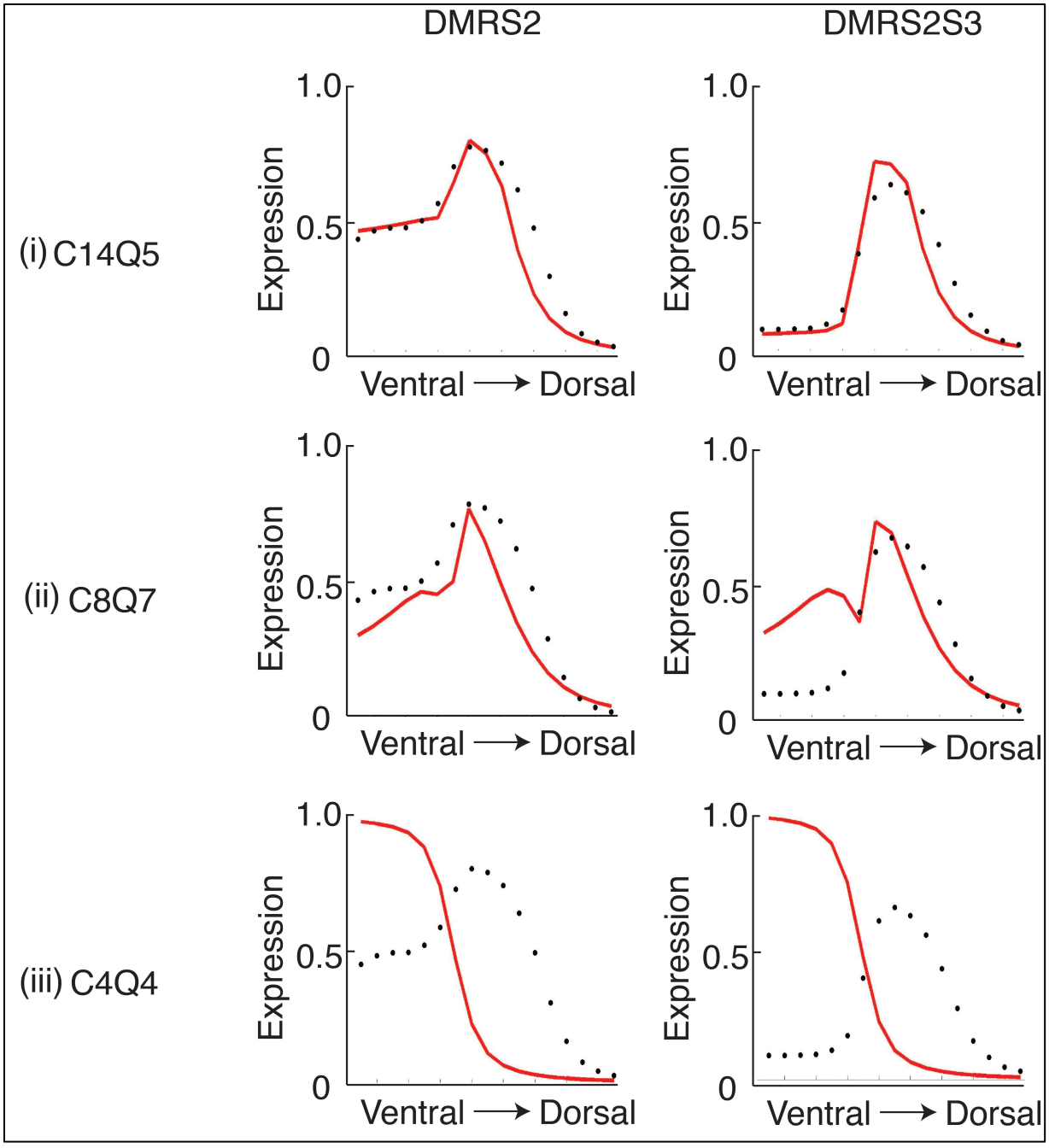


Figure 3.4 cont'd



The performance of the models was clearly co-dependent on both the cooperativity and repression formulations. Overall cooperativity model performance varied according to the type of quenching model, and the best quenching scheme for one cooperativity scheme was not necessarily optimal for other cooperativity schemes. Such interactive effects are likely a reflection of parameter compensation, as we discuss below. Nonetheless, similarly to the cooperativity scheme trends, there were trends in overall performance of quenching schemes, as revealed by the appearance of columns. In one scheme, we did not estimate parameters for repression, but set quenching values for Snail exactly to those identified for another short-range repressor, Giant (*20*). This approach was the least successful, as demonstrated by the overall higher RMSE values in column Q1. We have arranged the display of cooperativity and quenching schemes in figure 3.4A so that models with fewest parameters are located in the top left corner, and those with the most in the bottom right corner. Models with more parameters tended to outperform those with the fewest, as expected, but it was notable that this was not a strict correlation; the models with the most (35) parameters were not as effective as those with fewer. Additionally, there were measureable differences between models with identical numbers of parameters, suggesting that the different formulation of the schemes was interrogating aspects of enhancer grammar critical for the *rho* enhancer variants we were fitting. Overall, top performance for fitting our data set was seen for a model incorporating cooperativity values parameterized in three “bins” of 60 bp (C14) and quenching in four small 25 or 35 bp bins (Q5, Q6). As we discuss below, it is not wise to simply select one or two models based on this one aspect of performance,

because top scorers in this global analysis can change depending on starting conditions.

Overall RMSE provides a measure of the global performance of each model, but it is possible that models of similar RMSE have widely disparate performance on certain constructs. To examine this issue, we plotted a heat map illustrating model performance on a construct-by-construct basis (Fig. 3.4B). The models, represented by rows, were ranked from best to worst based on global performance, and individual fits for each of the 38 *rho* enhancer variants were plotted in columns. Several clear trends emerge from inspection of this figure; first, the higher ranked models (blue rows) have generally lower RMSEs across the board, as expected. Second, groups of models with similar structure had similar weaknesses on particular constructs. The top 35 ranked models assign separate parameters for homotypic interactions for Dorsal, Twist, and Snail; all of these had somewhat higher RMSEs for constructs 4, 5, and 17 which are the only *rho* variants in which Twist alone is perturbed. We speculate that the benefits from decoupling Dorsal and Snail cooperativity in these models more than outweighs the weaker performance on these three constructs. Third, several columns stand out which represent constructs that were generally fit less well by most models. Surprisingly, this included the wild-type *rho* enhancer (column 1); on further inspection, constructs 25, 28, and 32 showed a similar trend; all of these constructs contain the wild-type activator ensemble, whereas most of the data set used for fitting represented mutant *rho* forms with weakened activators, and lower expression in the dorsal regions of the stripe, where Dorsal and Twist proteins are limiting. Examination of individual plots revealed that the error arises from the models underestimating the activator potential specifically

in these regions of the embryo (Fig. 3.4C). The lower quality fit for these constructs does not represent a failure of the modeling; rather, these constructs are likely to be especially informative for activator function, and a modeling effort that entirely lacks these types of constructs would be more over-fit and less informative than the present one. Construct 17, lacking both Twist activator sites, was also not fit as well as others; this construct represents the only test of Dorsal alone as an activator, presenting a sparsely represented part of the perturbation space. Construct 35 had the overall poorest fit, representing a substitution in the bHLH sites and removal of Dorsal 1. The expression of this gene was the widest of all observed patterns, and showed poor fitting specifically in the dorsal-most regions; we suspect that the perturbation had multiple effects as discussed above.

Examination of individual plots for specific genes provides further insight into the nature of which features influence RMSE scores the most (Fig. 3.4C). A top-ranking model (C14Q5) accurately captures high and low levels of Snail repression accurately, falling down only in underestimating the expression of *rho* in regions most limiting for Dorsal and Twist. An intermediate-scoring model (C8Q7) was partially successful in capturing general trends of activation with occasional overestimation of activity. This model had some problems with Snail repression as well. The lowest performing model (C4Q4) suffered from multiple misfits in Dorsal/Twist activity, as well as a general absence of repression by Snail in the mesoderm. Thus, examination of model performance at three levels – global RMSE, construct-by-construct RMSE, and specific portions of the expression patterns - provides complementary insights into the nature of

how the training data are fit, and the potential utility of the models on other enhancer sequences.

Parameter values provide biochemical insights into enhancer function

Parameter estimation is a critical process in mathematical modeling of any process of scientific interest. The absolute as well as relative values of different parameters can provide us unique molecular insights into regulatory processes mediated by activator and repressor proteins on enhancers. Theoretically, any number of parameters can be assigned in order to describe a process in mathematical terms. However, larger number of parameters requires larger datasets to accurately estimate parameters and avoid over-fitting.

Scaling Factor –

Although we allowed scaling factors to assume any positive value, for all 120 models, Dorsal scaling factor values ranged between 2.75 e^{-18} and 0.0489, although extremely low values were observed for relatively few models. For top 10 models, average Dorsal scaling factor value was 0.0049 (Figure 3.5A(i)). Twist scaling factor was estimated in the range of $1.35e^{-20}$-1.796. For top 10 models, average Twist scaling factor value was 0.0189. Twist scaling factor values were observed to be more model-dependent than Dorsal, making the comparisons between Dorsal and Twist scaling factors less reliable. Snail scaling factor came out to be quite consistently in the range of values of around 40-50 for top 40 models (mean value 42.57±8.59). A few models estimated very high values, but these were ranked at bottom (Figure 3.5A(iii)).

When we arranged scaling factors by model performance, certain trends became apparent. Dorsal scaling factor values clustered into three groups. First cluster

comprised of models ranked among top 40. Most of these models had lower values of Dorsal scaling factor than the second cluster, which consisted of models ranked from 40 to 90. The last group had higher variability than first two groups. Lower Dorsal scaling factor values for better performing models suggests that Dorsal has a weak activation potential. For Twist, we found more variability in the spread of values than Dorsal (Fig. 3.5A(ii)). Lack of consistency among closely ranked models suggests that this parameter does not have a strong influence on model performance, which is also corroborated with the fact that the enhancer has only two Twist sites, and consequently, it does not constrain the model strongly due to lesser number of terms in model formulation. Snail scaling factor values appeared to cluster into two groups. First cluster consisted of values from models ranked among top 40. As mentioned earlier, there was a strong tendency among these models to have Snail scaling factor between 30 and 50. Models ranked 40 to 120 had high variability, making any judgment about the parameter values inconclusive. The tendency of top 40 models to have high repressor scaling factor values indicates that Snail has a strong repressive influence on gene expression, much higher than activation potential of the two activators.

Scaling factor values might be expected to show a compensatory relationship with cooperativity or quenching. Compensation between scaling factor and cooperativity is indeed observed in some cases, such that parameter sets with similar RMSEs can be found where lower Twist scaling activity is compensated by higher Dorsal-Twist cooperativity (Fig. 3.5D). In most cases, however, parameter sets identified with similar overall RMSE tended to have similar scaling factor values for all three proteins. This observation is useful if we want to derive meaning from parameter values.

Cooperativity – We estimated cooperativity between different proteins in two different ways – either for homotypic and heterotypic interactions, or on a protein-by-protein basis. In the latter case, we included homotypic cooperativity between Dorsal, Twist and Snail, and heterotypic cooperativity between Dorsal and Twist. In global ranking of models, the former models, where cooperativity is either homotypic or heterotypic, performed poorly. Top 35 models were exclusively populated by models with cooperativity estimated separately for different proteins.

In our analysis of all models with heterotypic and homotypic cooperativity, we observed that heterotypic cooperativity has frequently high positive values (in 39 out of 64 model combinations) as compared to very low positive values (16 out of 64; 9 models ambiguous). These high positive values of heterotypic cooperativity reflect cooperativity between Dorsal and Twist, which is also known from previous studies (*57*). Homotypic cooperativity values were found to be frequently very low (36 out of 64) than very high (12 out of 64). In formulation of thermodynamic models, very low values for cooperativity would effectively reduce the statistical weight of the state with two cooperating proteins bound, reducing the contribution of this state to gene expression. Homotypic cooperativity here refers to cooperativity between similar sites for both activators and repressors. As discussed below, we get highly divergent values for cooperativity between activators and repressors, which indicate different mechanisms for these two categories of proteins, so negative cooperativity values for three different proteins estimated together may be difficult to interpret.

In models with separate parameters for different proteins, Dorsal-Dorsal cooperativity was observed to be positive for all 48 model combinations except one (Fig.

3.5B(i)). The most frequently observed trend was positive cooperativity, which remained constant with increasing distance. Twist-Twist cooperativity showed a similar trend in 47 out of 48 model combinations, with slightly larger cooperativity values than Dorsal-Dorsal cooperativity (Fig. 3.5B(ii)). These distance-invariant cooperativity values imply that the cooperativity is not required for binding to DNA and may not require linkage of binding sites. This is a novel aspect of cooperativity between these factors and should be explored experimentally. In contrast with these positive cooperativity values, Snail-Snail cooperativity was almost always very low (45 out of 48 combinations, Fig. 3.5B(iii)). These extremely low cooperativity values would undermine the effect of binding of two Snail sites on repression, with most of the repression being estimated using states with single sites bound. Dorsal-Twist cooperativity was almost always found to be very high (45 out of 48 models), which is also in agreement with heterotypic cooperativity estimated for our other set of models (Fig. 3.5C(i)).

We hypothesize that low cooperativity between repressor sites and high cooperativity between activator sites has implications for enhancer structure and how it may evolve. Low cooperativity between Snail sites, along with the fact that it is a short-range repressor, will require it to be positioned at multiple sites in an enhancer for effective repression. On the other hand, low scaling factor and high cooperativity between activators implicates the importance of multiple activator sites to mediate transcriptional activation.

Quenching

Quenching of Dorsal activator function by Snail – Snail is a short-range repressor, but details of its range of action are lacking. Our analysis of snail repression function on

Dorsal revealed a distance-dependent decrease in repression efficiency of Snail. This function was observed for 50 out of 120 model combinations (Figure 3.5C(ii)), and was present in 15 of top 30 models. We observed this trend for quenching models in combination with all different cooperativity models.

Quenching of Twist activator function by Snail – The distance-dependent decrease of repression efficiency by Snail was observed in case of Twist for a minority of models (7 out of top 30). In a small number of cases (11 out of 120), Snail quenching on Twist was observed to be either low for proximal bins and higher for intermediate bins (Figure 3.5C(iii)), or near zero for all distances. This difference may imply that Snail acts differently on these two activators. The inconsistency among quenching trends supports our earlier observations on Twist scaling factor and our hypothesis that Twist plays a smaller role in enhancer function than Dorsal. Additionally, one of Twist sites is compromised by competition for binding by Snail. Taken together, these results suggest that although activation in *rho*NEE is carried out mediated by Dorsal and Twist together, Snail preferentially targets Dorsal to interfere with activation in mesoderm.

Quenching function from Fakhouri *et al.*– We noticed different trends in cooperativity when we compared how this particular quenching function influences cooperativity values. In general, we observed that trends for cooperativity parameters with this quenching function were similar to those observed in combination with quenching functions that were not fixed from the outset. Therefore, we think that this pre-determined quenching function did not constrain the cooperativity parameters significantly.

Thermodynamic models depend on information about the sequence of DNA sequences in question, the levels of regulatory proteins acting on the DNA, and the known DNA interaction potential for the proteins, represented by position weight matrix information. We refit all 120 models on our dataset using alternative descriptions of the Dorsal, Twist and Snail PWMs obtained by high-throughput analyses (SELEX and bacterial one-hybrid). The global performance of the models was overall worse, resulting in higher RMSEs (Supp. Fig. S3.2). As we show below, this is largely a function of the inferior performance of the Snail PWM. Construct 32, which explores the activity of just the Snail 2 and Snail 4 sites, stands out as an egregiously poorly fit construct for all models, with problems not just with activation in dorsal regions, but also a strong overestimate of repression in ventral regions. Tellingly, the SELEX PWM used for this run fails to predict Snail 3, which experimental evidence indicates is strongly contributing to repression activity, and it misses the mark as well on Snail motifs 1 and 4, therefore the models are presumably mostly fitting parameters to the remaining Snail 2 site, and overestimating the capacity of this single motif to influence repression in the mesoderm. In addition to these differences, however, there were also strong similarities between the two model-fitting results: C1 and Q1 remained the weakest schemes for cooperativity and quenching, and models with individual parameters for each homotypic cooperative interaction were generally stronger performers.

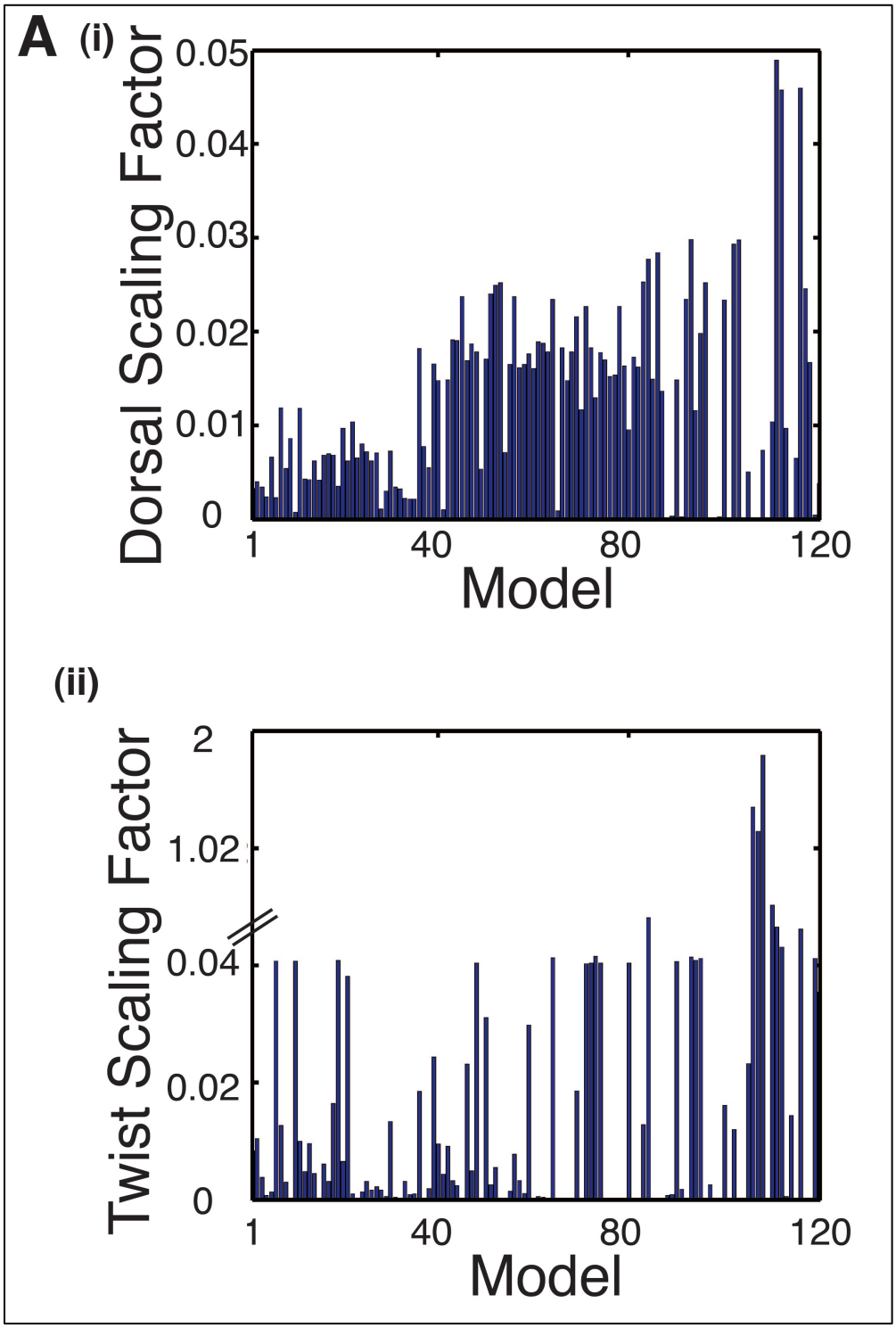


Figure 3.5. Parameter values derived from 120 models. (A) Scaling factors for Dorsal (i), Twist (ii) and Snail (iii) obtained for 120 models, arranged according to model rank on X-axis. (B) Homotypic cooperativity trends observed in protein-binned models.

Figure 3.5 cont'd

The plots show log parameter values obtained for 5 runs in different bins for (i) Dorsal-Dorsal, (ii) Twist-Twist and (iii) Snail-Snail cooperativity. Run 2 had high RMSE for this model and is depicted as a dotted gray line. (C) (i) Heterotypic cooperativity between Dorsal and Twist estimated by one of the protein-binned models. (ii) Quenching trends observed for Snail on Dorsal and (iii) Snail on Twist. (D) An example of parameter compensation observed. (i) Twist scaling factor was lower for one of the runs of CMAES (arrow). (ii) Heterptypic cooperativity was higher than other runs in all distance bins (arrows). (iii) Snail quenching on Twist was higher (blue circle). (iv) All runs had similar RMSE.

Figure 3.5 cont'd

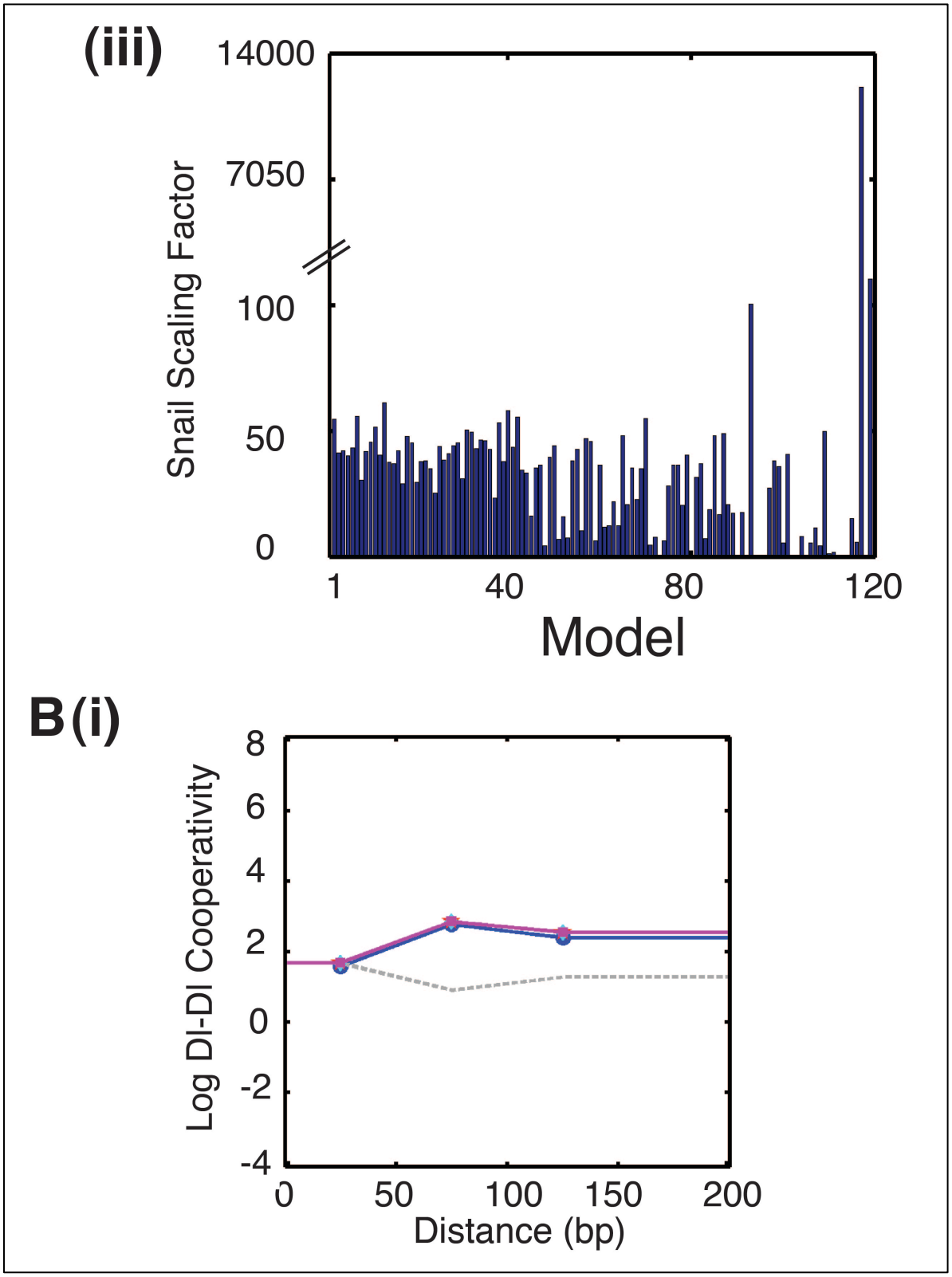


Figure 3.5 cont'd

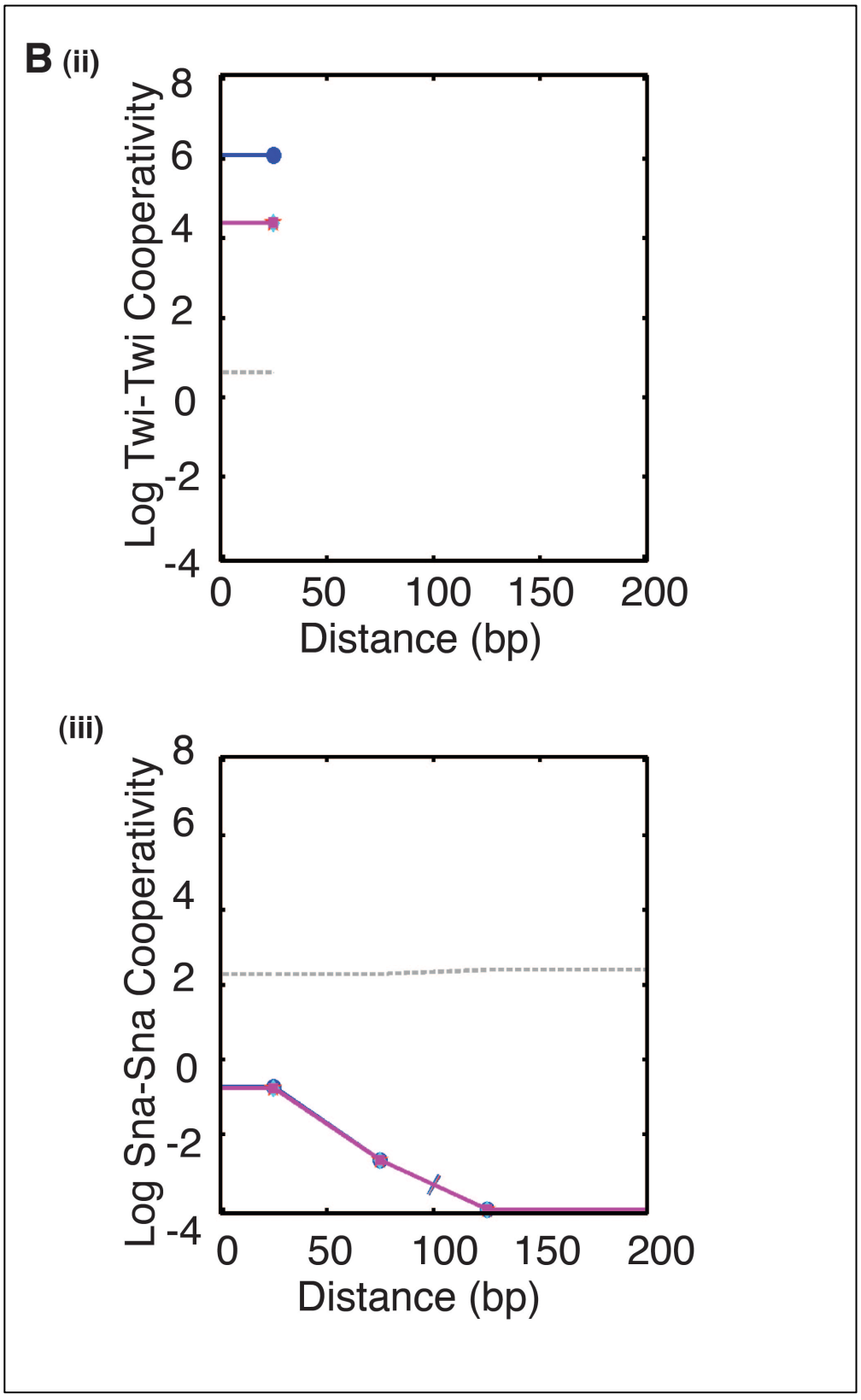


Figure 3.5 cont'd

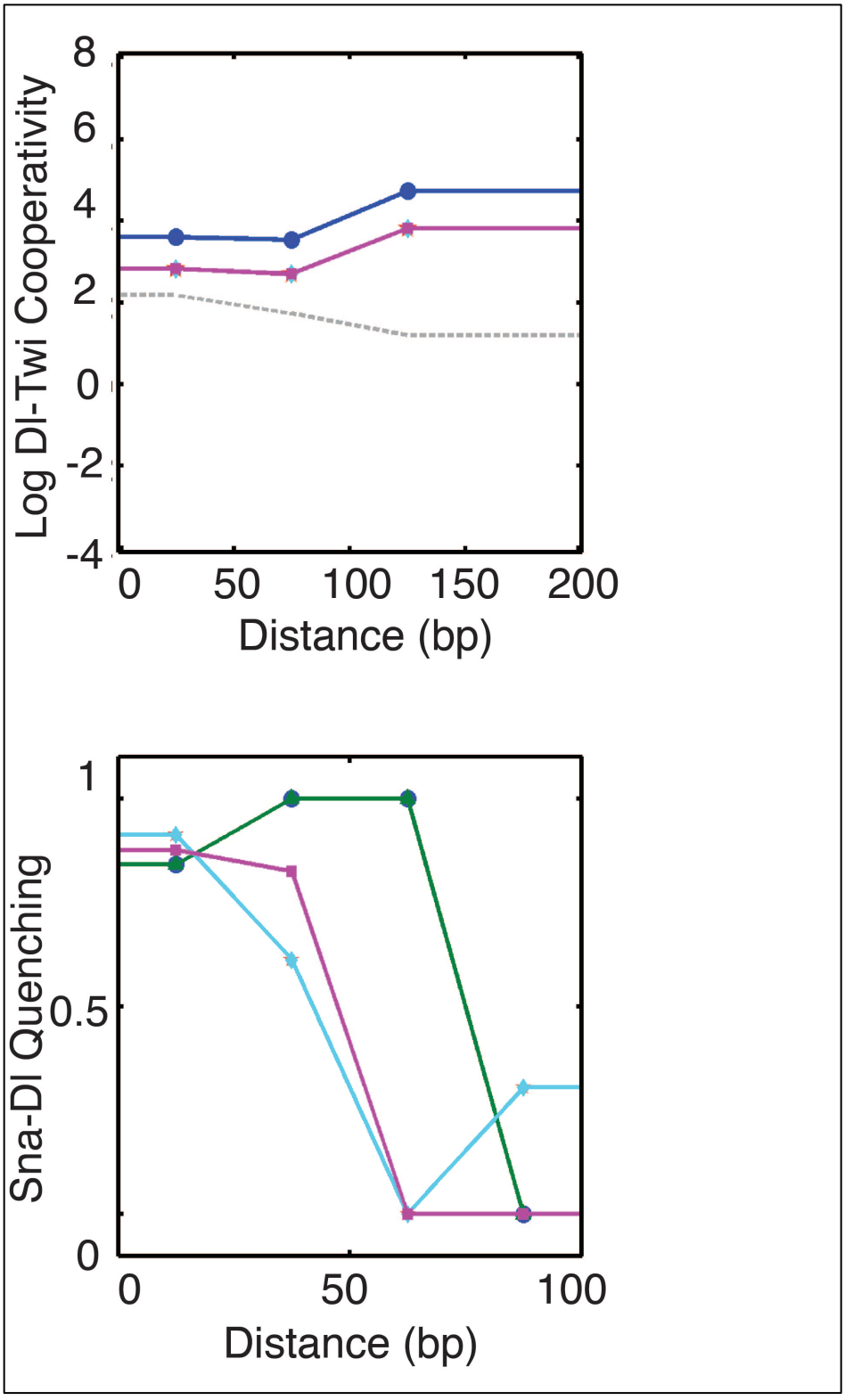


Figure 3.5 cont'd

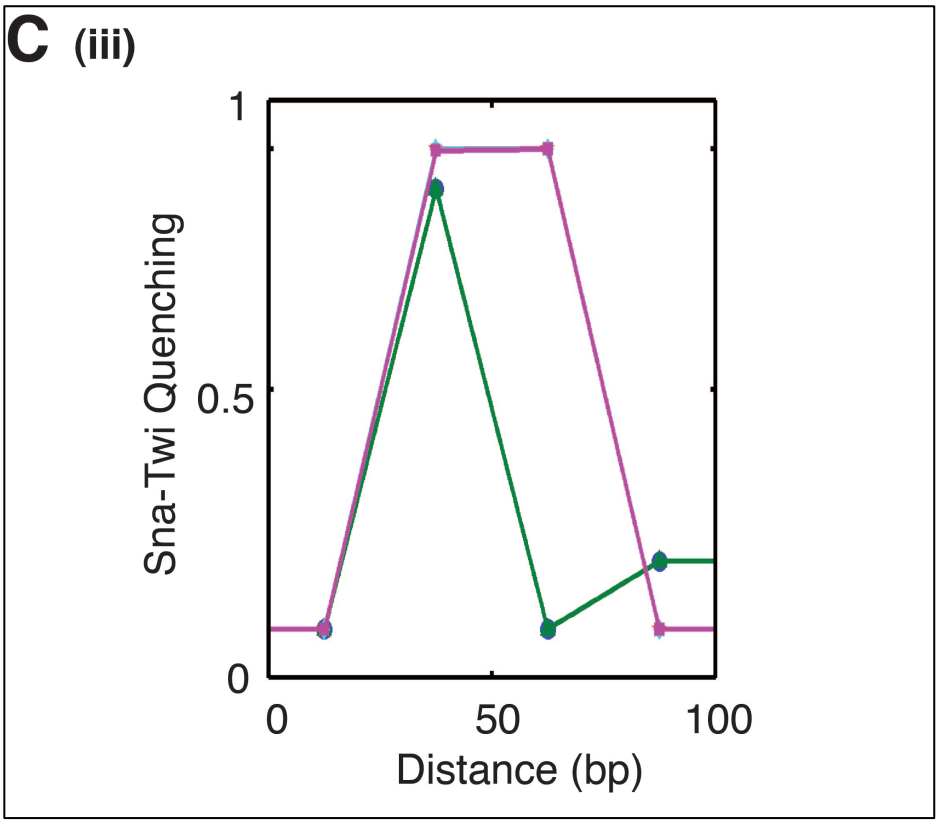
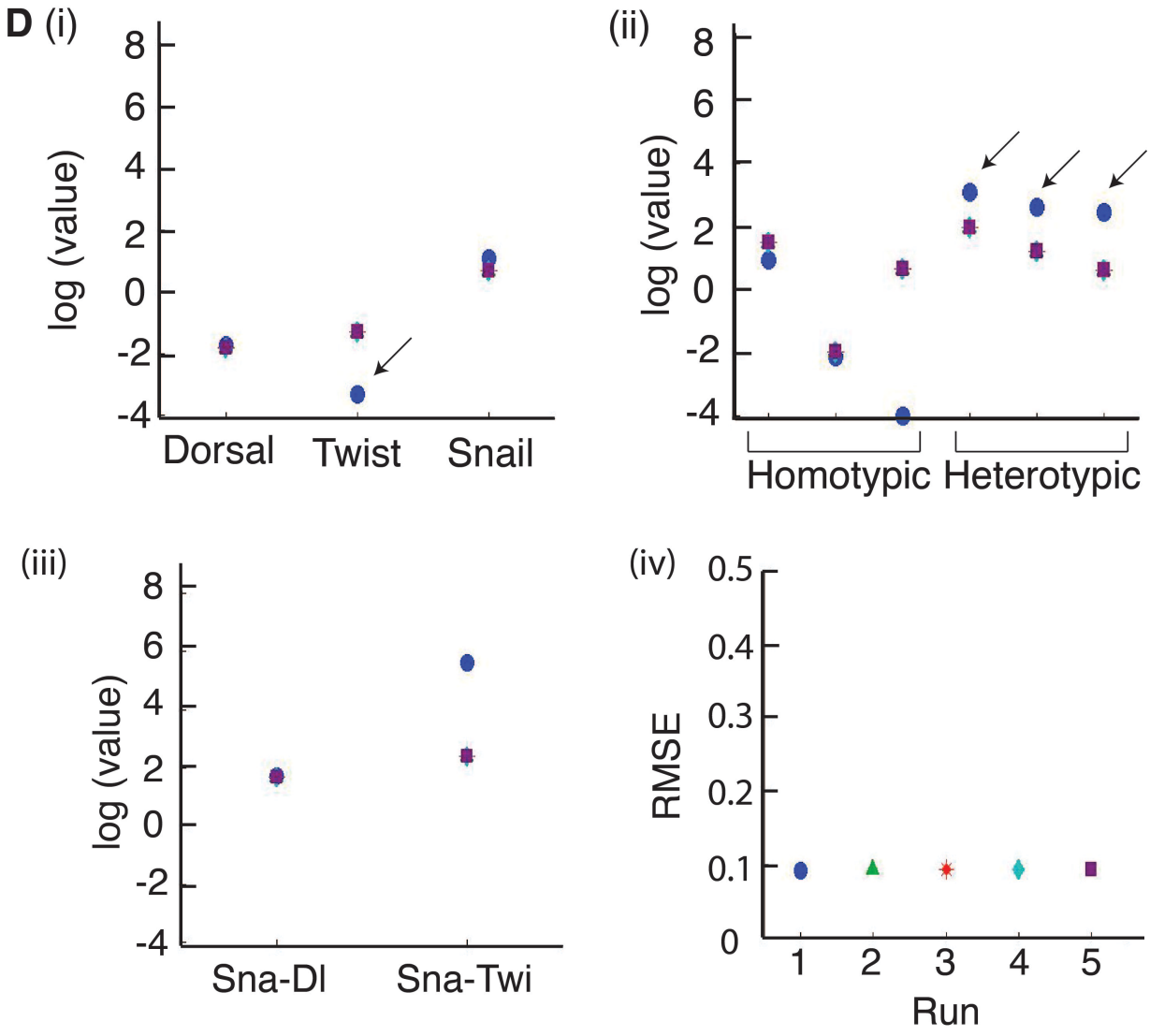


Figure 3.5 cont'd



Effect of changing PWMs and thresholds on model performance

To more comprehensively explore the impact of PWM values on model fitting, we tested combinations of all available PWMs for Dorsal, Twist and Snail (3 PWMs each) which were derived from different experiments (see Materials and Methods), and used them to define the landscape of binding sites on the *rho* enhancer. All possible combinations of PWMs (27 total) were evaluated using 24 models, representing many of the better scoring models, as well as some of the weaker performers, and re-estimated parameters for each. The global performance for each model as a function of differing PWMs was plotted as a heat map and columns were arranged so that the PWM settings that generated overall the lowest RMSE values are at the left, and overall best performing models ranked in the rows toward the top (Fig. 3.6). A striking correlation between the Snail motif and overall performance divides the field into three domains; eight of the top nine scoring PWM settings included the Snail PWM derived from footprinted binding sites (rather than SELEX data), which captures the four known Snail binding sites on the *rho* enhancer (Fig. S3.3). The definition of the Twist and Dorsal sites had less of an effect, as there was no clear correlation between overall performance of the 24 models and the choice of these PWMs. In the case of Twist, all three identified motifs are similar in composition, and make similar predictions with respect to binding sites in the enhancer, so this finding was not surprising. For Dorsal, the PWMs differ more, but they still identify similar sites, and thus the performance is not dramatically altered. The conclusion from this analysis is that uncertainty in the information at the levels of PWM can globally alter the power of thermodynamic models to correctly fit a modeling set, and presumably predict novel enhancer sequences. Regardless of this influence, top-ranking models as a group appear to perform

consistently better than lower ranking ones, although for a particular choice of PWMs the individual performance can move a model up or down in the ranking. This finding points to the utility of testing groups of related models for optimal predictive power, similar to the conclusions for using motif search algorithms in bioinformatic studies (*60*).

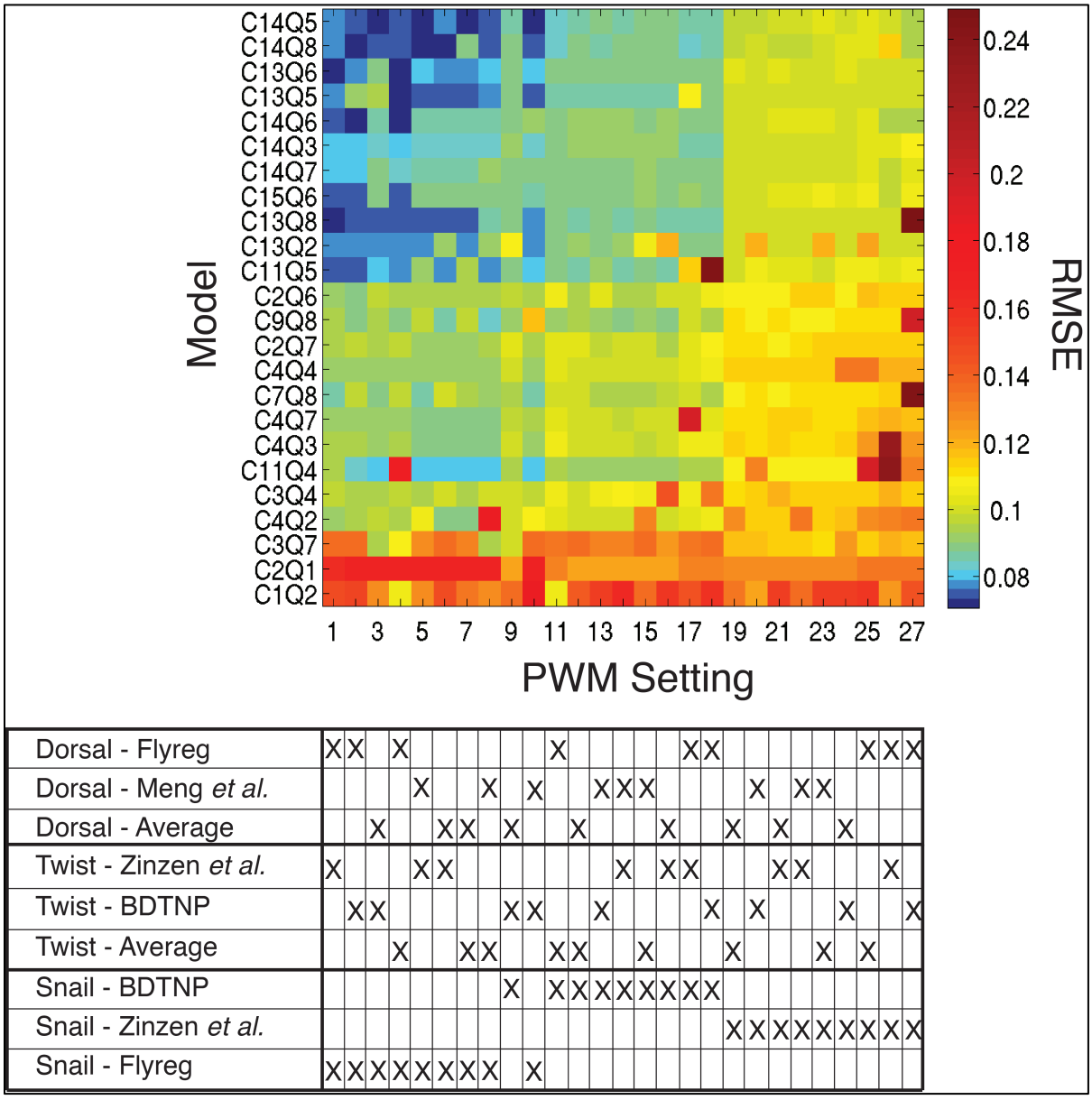


Figure 3.6. Effect of different PWM settings on model performance. The heatmap shows performance of 24 selected models on the dataset using different PWMs for Dorsal, Twist and Snail to annotate binding sites. Columns of PWM settings are arranged with lowest global RMSE values to the left. Rows of models are arranged with lowest global RMSE towards the top. PWMs used for different settings are denoted by cross marks in the lower panel (see Fig. S3.4).

Cross-validation Analyses

The structure of dataset can have dramatic influence on model learning and performance. Cross-validation involves testing model performance on subsets of the data. Models which perform poorly during cross-validation indicate the propensity of these models to overfit rather than learn from a given dataset. For the 24 models, we performed cross-validation using two different strategies – i) leaving parts of dataset out based on specific groups (all constructs with Dorsal site mutations, Snail site mutations etc.), and ii) taking out random parts of dataset. In general, model performance worsened more when specific sets were left out than random leave-sets-out (Fig. 3.7). This observation has two implications – first, it suggests that specific sets of constructs contribute significantly to model learning, and the different sets of constructs are informative to different models. Second, it suggests that our dataset is large enough so that many models show robustness to minor decreases in number of data-points, meaning that the number of constructs is large enough for model training. Specific sets cross-validation tests whether models are learning about contribution of different classes of proteins to enhancer function from different parts of dataset. This cross-validation indicated that most models showed deterioration of performance when specific sets were left out, indicating the importance of careful design of constructs to learn about specific features of enhancer function. It is interesting to note that the above observations were not shared by many poor-performing models. Several of these (e.g., C3Q4, C3Q7, C1Q2 and C2Q1) showed drop in performance during both random as well as specific-set cross-validation, which suggests that the parameters estimated by these models are probably overfit and are not robust enough to small changes to

decrease in data, and that these models may not perform so well on predicting gene expression from enhancers outside of model training.

For model C4Q4, which ranked 120, model performance improved (RMSE decreased) during systematic and random cross-validation, more so in latter. Specific-set cross-validation suggested that model had high predictive power when activator and repressor sets were left out. Random cross-validation also revealed that model performance improved 2-3 times when data was sliced into different parts randomly. However, we observed large changes in model performance from run to run (data not shown), which was peculiar to this particular model. This increased variability may be the reason behind this apparent improvement in performance.

When we analyzed how leaving out different sets of constructs out of model training affects model performance, we observed that, multiple sets of constructs showed increase in RMSE (poorer predictive power) when these were left out. 6 models had higher RMSE for each and every set was left out. Snail constructs had the maximum increase in RMSE (poorest predictive power) when left out, in 15 out of 24 models.

We also analyzed how leaving specific sets of constructs affects model predictions on those specific constructs. For example, leaving repressor site constructs out of parameter estimation may lead to worse predictions on these constructs. For bHLH constructs, performance of model C11Q4 worsened across the board. We think this decrease in performance may have been caused by the estimation algorithm getting stuck in a local minimum. Only a few models showed decrease in performance for prediction on bHLH constructs – C13Q6, C13Q5 and C11Q5. In case of Dorsal site

constructs, performance generally worsened for almost all models for construct 3 (DMRD2), construct 8 (DMRD1D2). Additionally, model performance worsened for only a few top and bottom models for constructs 11 (DMRD1D3) and 12 (DMRD1D4). Lastly, large drops in model performance were observed for constructs 15 (DMRD2D3) and 16 (DMRD2D4). For constructs with Dorsal and Twist sites mutated together, models C13Q6, C13Q5 and C11Q5 again (as in the case of bHLH constructs) showed worse predictions on constructs 9 (DMRD1T1) and 10 (DMRD1T2), and almost all models had worse performance on construct 14 (DMRD2T2), 18 (DMRT1D3) and 19 (DMRT1D4).

For Snail constructs, model C11Q4 has worse performance across the board, almost all models had overall large increase in RMSE for all Snail constructs. For Twist constructs, model performance worsened for all top models until C2Q6 (Rank 43) for all three constructs left out. Also, model performance was worse for construct 17 in general (DMRT1T2). Collectively, these differential sensitivities of different models to different constructs imply that these models are learning important features about the effects of different proteins, which consequently decrease the predictive power of the models when these data points are excluded from parameter estimation.

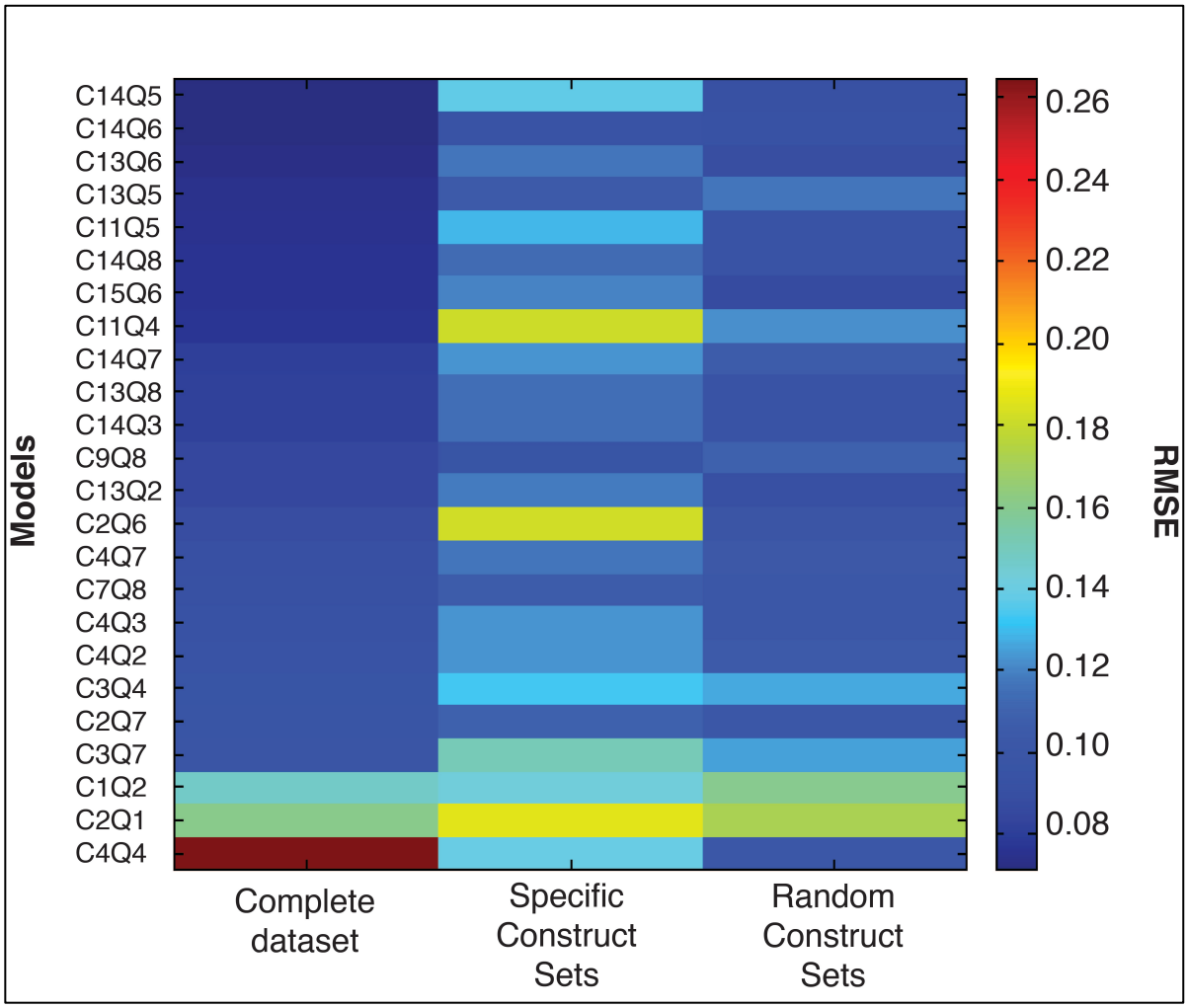


Figure 3.7. Effect of cross-validation on model performance. The heatmap shows performance of selected 24 models on complete dataset of 38 constructs (left column), performance when specific construct sets are left out (middle column), and when random sets are left out of parameter estimation (right column). The colors denote different RMSE values (color bar on right).

Sensitivity Analysis

In our analysis of scaling factors, top-ranked models showed only second-order relative sensitivity for Dorsal. For Twist, among the ten models which ranked among top thirty (out of 120 total), 9 showed some first order sensitivity. Snail showed the highest overall sensitivity among the scaling factor parameters. For all 7 models ranking in top 10 (out of 120), Snail had lower second order sensitivity than first order. Among the 24 models, 23 models had some first order sensitivity.

In case of protein-binned models, Dorsal-Dorsal cooperativity showed different levels of sensitivities. Generally, this parameter showed low first order sensitivity and high second order sensitivity. Twist-Twist cooperativity showed either very low first order sensitivity, or completely second order sensitivity. Snail-Snail cooperativity also showed different first and second order sensitivities in different distance bins. In general, all models analyzed showed high second order sensitivity. Dorsal-Twist cooperativity showed different sensitivities for different bins. This parameter showed some, albeit very low first order sensitivity, and high second order sensitivity. A few bins also showed completely second order sensitivity. For binned models, heterotypic cooperativity showed relatively low first order sensitivity, and high second order sensitivity. In a few cases, it showed completely second order sensitivity for a few bins in different models. Homotypic cooperativity showed low first order sensitivity and high second order sensitivity.

Only a third of models showed high first order sensitivity for Snail quenching parameter for Dorsal. In binned models, some bins showed higher first order sensitivity for this parameter than others, usually the ones for larger bins. The parameter for Snail

quenching on Twist exhibited low first order sensitivity, and high second order sensitivity. In general, continuous quenching functions showed some first order sensitivity, and low second order sensitivity.

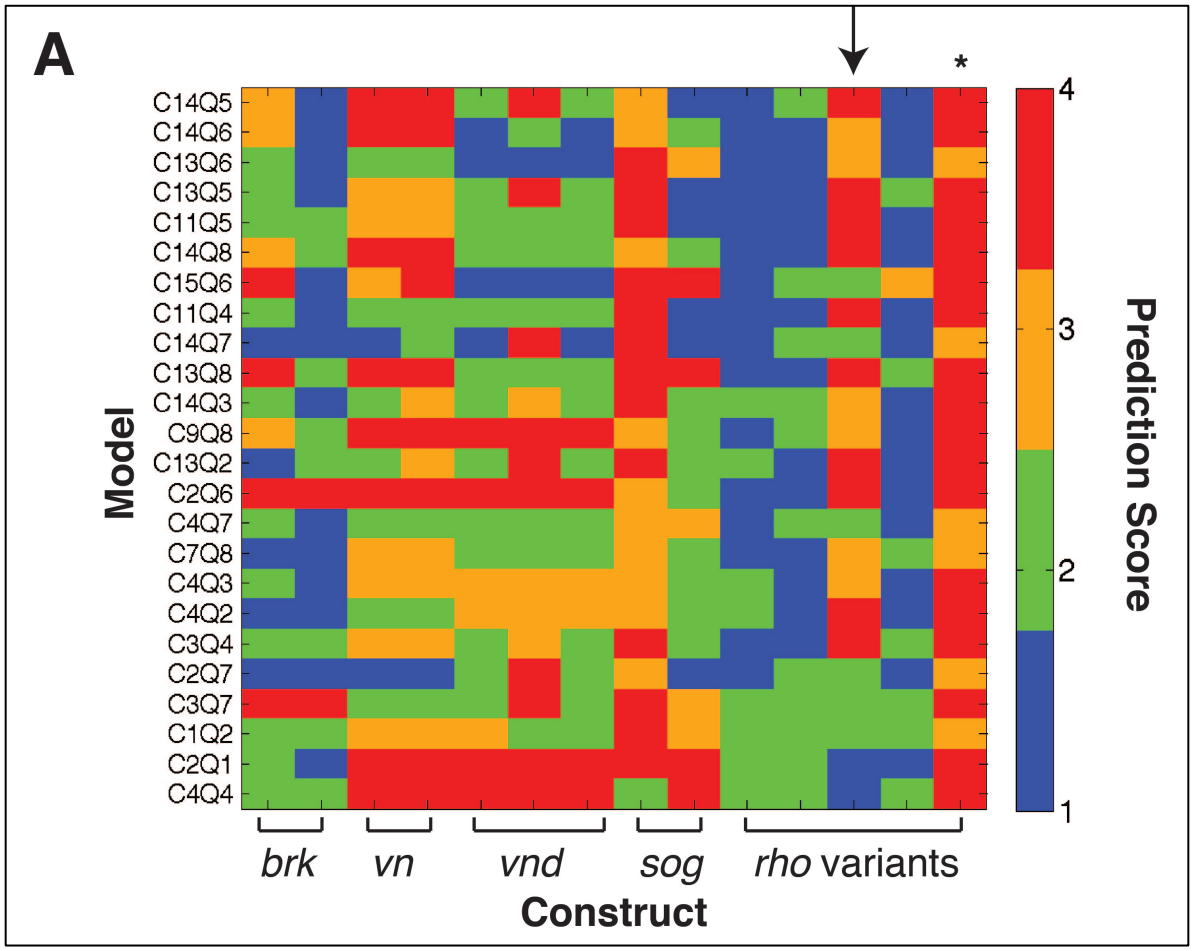


Figure 3.8. Prediction of gene expression from other dorsal-ventral patterning enhancers. Prediction of (A) correct activation in neurectoderm and (C) mesodermal repression by Snail for 14 different enhancers by 24 models. Prediction of (B) activation by Dorsal and Twist in mesoderm and (D) mesodermal repression by Snail for 7 different enhancers by 24 models. See materials and methods for scoring of expression patterns.

Figure 3.8 cont'd

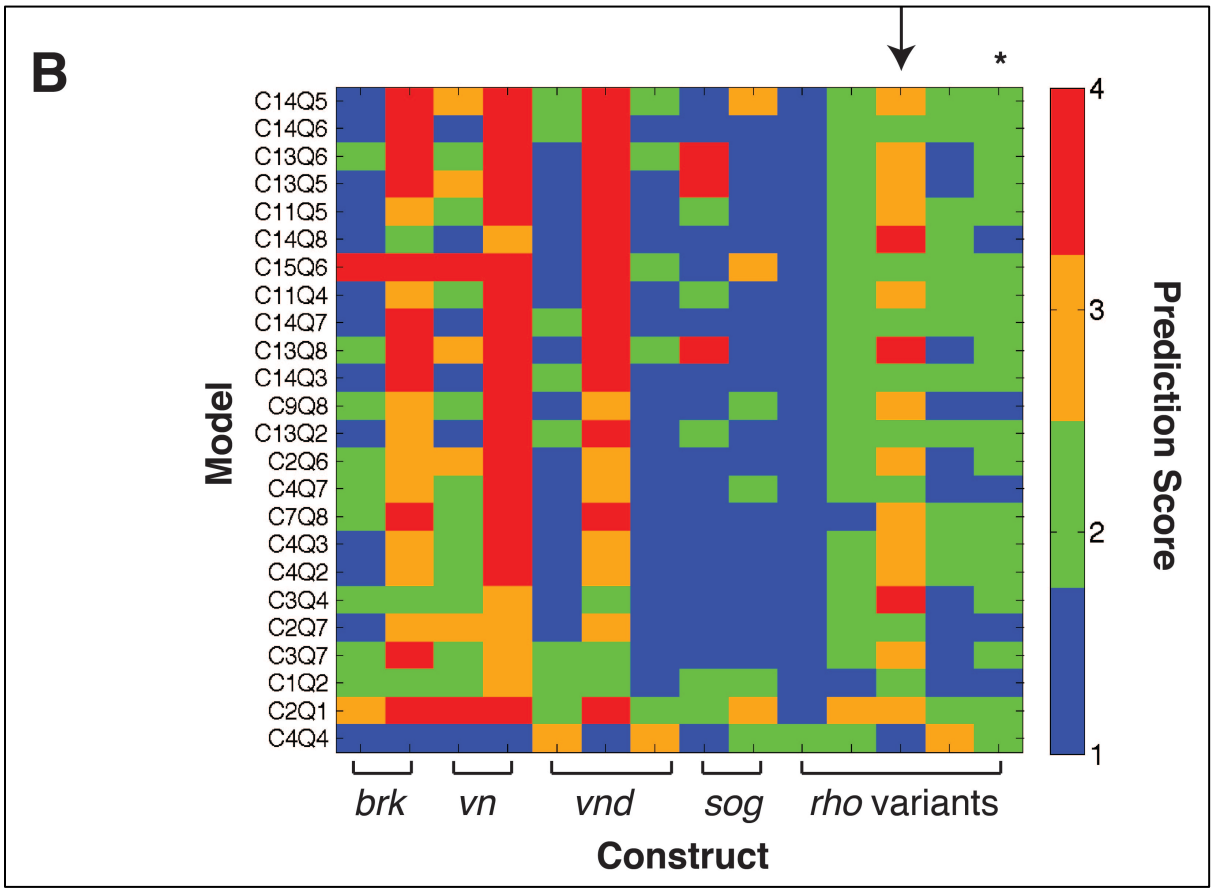


Figure 3.8 cont'd

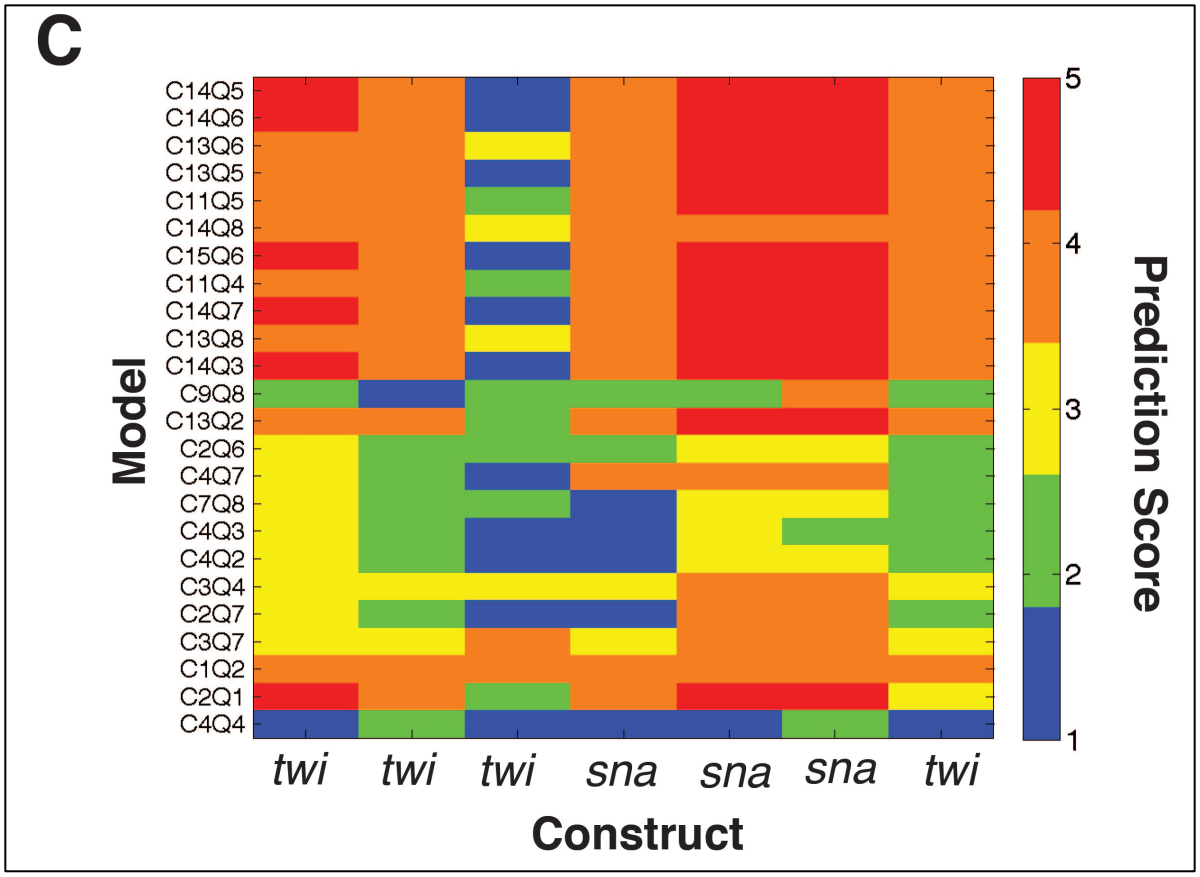
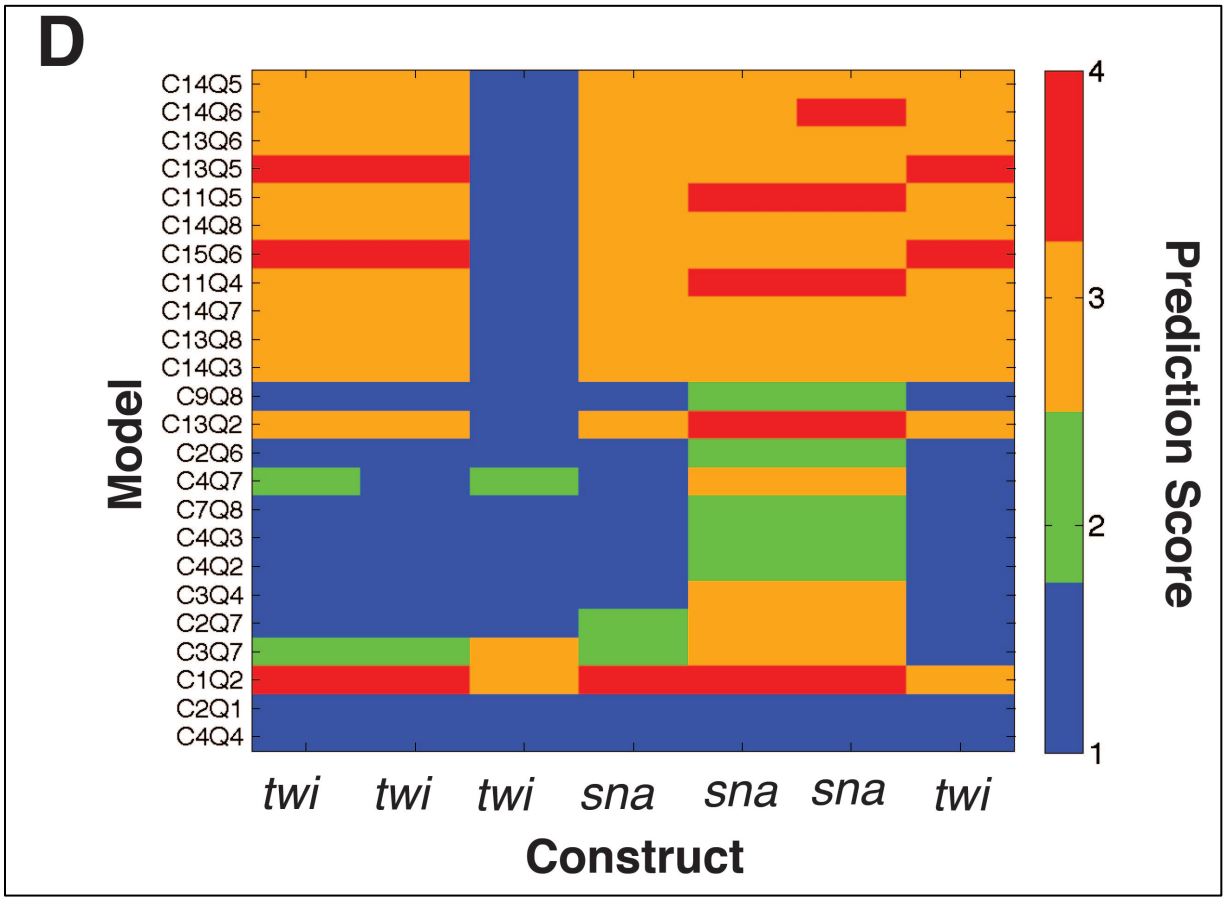


Figure 3.8 cont'd



Thermodynamic models can quantitatively predict gene expression from other dorsal-ventral patterning enhancers

In order to test the parameters obtained from our analysis of the *rho* enhancer, we used them to predict the expression patterns for other enhancers for genes also regulated by Dorsal, Twist and Snail. These enhancers represented the variety of genes expressed in both mesoderm and neurectoderm. Fourteen enhancers active in neurectoderm and seven enhancers active in mesoderm were cloned and expression data was obtained for these twenty-one enhancers in the same manner as for *rho* enhancer variants (Fig. S3.1).

For this analysis, we tested the twenty-four models previously analyzed for sensitivity and global performance as described above. For enhancers active in neurectoderm, we scored predictions based on activation in neurectoderm and repression in mesoderm. The top eight models selected from this group (which scored within the top 14 of 120 models) predicted correctly the activation and Snail repression for most of these enhancers (Fig. 3.8, panels A and B). For some enhancers, models were unable to capture one of the features. For example, we tested two constructs for *brinker* (*brk*) gene, second construct being the larger enhancer fragment. Top models can capture the repression for shorter enhancer while giving erroneous predictions on activation. The predictions for larger enhancer are exactly opposite (compare columns 1 and 2 in Fig. 3.8, panels A and B). Most of the bottom ranked models cannot predict enhancer activity, and their performance is especially poor regarding prediction of Snail activity.

This dataset, which was not used for parameter estimation, also included five putative rhomboid enhancers derived from evolutionarily distant fly species – *D. erecta, D. ananassae, D. mojavensis, D. grimshawi and D. virilis*. These species represent a wide spectrum of evolutionary divergence time from reference species, *D. melanogaster*. These putative enhancers drove gene expression in neurectoderm, although the expression level for one of the enhancers, putative *D. mojavensis rho*NEE, was slightly lower than other NEEs. Strikingly, many of our top models were able to predict gene expression from these divergent enhancers ("*rho* variant constructs" in Fig. 3.8, panels A and B). However, almost all of the models were not able to capture the lower activation and Snail repression for *D. mojavensis rho*NEE and Snail repression for *D. virilis rho*NEE. The ability of several models to correctly predict expression indicates a conserved grammar for these proteins on these enhancers, while the inability to predict the crucial features for some of these enhancers suggests two possibilities – first, that the structures of our models cannot capture the grammar for some enhancers due to bin sizes or functions not tested in modeling, and second, that other factors, for example, differential nucleosome positioning are playing important roles in regulating expression for these enhancers that are not captured by sequence-based gene expression predictions.

Top-ranked models generally performed better for predictions on dataset. We also observed that performance on predictions was not highly correlated with performance on training dataset.

It is also worth noting that top 7 models perform differently on this ensemble of enhancers. Each model has different predictive power on different enhancers. We think

that this difference in performance is due to their diverse structures. Collectively, most of the top 7 models can capture the essential attributes of tissue-specific gene activation and repression for these enhancers.

We also analyzed seven enhancers from *twi* and *sna* genes, which drove expression specifically in mesoderm. Analysis of how different models predicted activation and repression for these enhancers revealed that mesodermal expression was hard to be captured by most of the models. Only one enhancer (#53, panels C and D in Figure 3.8) could be accurately predicted by our models. A closer look at the binding site organization of these enhancers revealed that most of the enhancers had numerous (9-17) dorsal sites, and construct 53 had only 3 Dorsal sites. The high dorsal-dorsal cooperativity estimated by most of the models resulted in high levels of expression predicted by models, often extending into neurectoderm. Nevertheless, the failure of different models to accurately predict expression for mesodermal enhancer suggests that more expression data from such enhancers is needed to capture the grammar for binding sites.

Top-ranked models can predict location as well as output of tissue-specific enhancers

Dorsal is the fly homolog of NF-kB and is the master regulator of dorsal-ventral patterning in early fly embryo. Dorsal binds to regulatory regions of around 100 genes and sets up the primary germ layers which lead to subsequent tissue differentiation. A variety of other sequence-specific DNA-binding proteins act together with Dorsal to activate and repress different genes. Tightly linked binding sites of Dorsal and Twist have been shown to be important for gene activation in neurectoderm, whereas binding

sites of strong and weak affinities are necessary for expression in neurectoderm and mesoderm respectively (*57*, *61*). Dorsal often acts together with Zelda, a ubiquitously distributed activator to regulate several target genes in neurectoderm and dorsal ectoderm (*62-64*). Several studies have shown how minor changes in grammar of binding sites for Dorsal and its interaction partners can lead to diverse regulatory outputs.

We utilized previously published data for genome-wide binding of Dorsal, Twist and Snail in early syncytial fly embryo and isolated the DNA sequences bound by these proteins to predict their functional output (*30*). Many of these have been shown to be bona fide regulatory sequences. We selected 75 bound regions to be used as input for our 24 selected models for predicting gene expression. These regions were chosen based on two criteria – firstly, the gene nearest to these regions is expressed in mesoderm or neurectoderm at this stage (2-4 hrs old embryo), and secondly, the expression pattern is known. We used publicly available community resources (BDGP and FlyExpress) to match genes with expression patterns, and scored the predictions from our models to the expected expression patterns (Fig. 3.9).

We used the same PWMs for our predictions that we used for parameter estimation. For some enhancers, we used lower thresholds to uncover weak binding sites, if we could not recover any sites using our original thresholds. These enhancers are widely divergent from our training dataset in terms of both number as well as arrangement of binding sites, an ideal litmus test to test thermodynamic models. We found that the top model can actually predict the gene expression in the correct tissue from these enhancer sequences (Fig. 3.9). Lower ranked models, e.g., C3Q7 and

C4Q4, cannot predict the activity of these enhancers correctly. Since many of the enhancers used for predictions have already been validated successfully in previous studies, we have a high confidence in assessments of these predictions. For genes with known expression patterns, our predictions predict that many of the genomic regions bound by transcription factors function as tissue-specific enhancers.

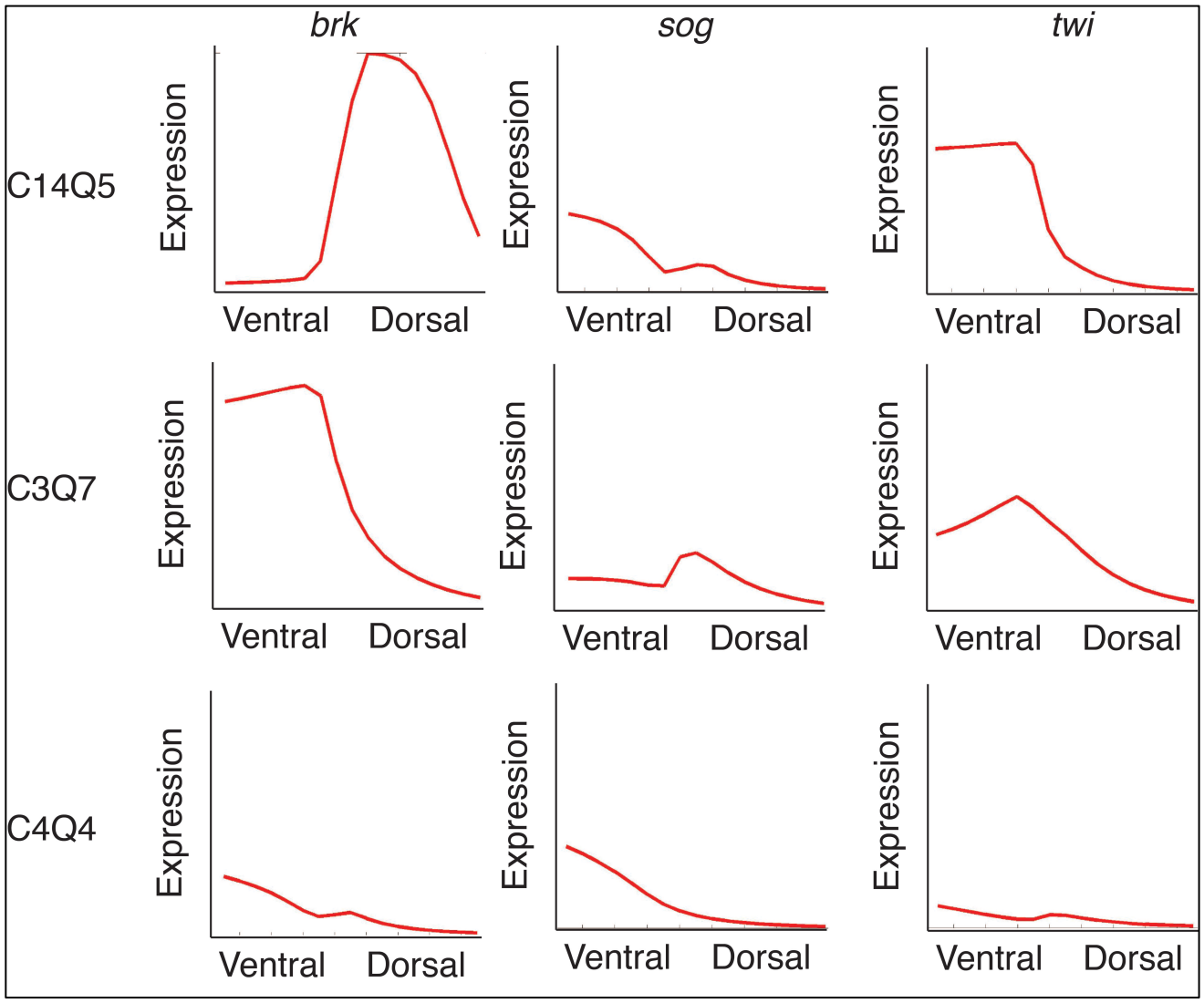


Figure 3.9. Prediction of gene expression from genome-wide ChIP-chip binding regions for Dorsal, Twist and Snail. Shown here are predictions for putative regulatory regions for *brk, twi* and *sog.* Top three panels show predictions for model C14Q5, middle three panels show predictions for model C3Q7 and bottom three panels show predictions for model C4Q4.

Discussion

Our study tests the ideas that sequence content of transcriptional enhancers is highly informative about function, and that the constraints on protein binding site arrangement reflect a common “grammar” for transcription factors that underlies their action on co-regulated cis elements. Both of these notions have experimental support, but their generalizability is not known, a key issue for the feasibility of mining the vast amount of genomic data representing population- and species-level variation. Recent studies have called into question whether indirect interactions between transcription factors and the DNA (via protein-protein interactions) may allow transcription factors to load promiscuously on regulatory sequences (*32*, *65*). If this “indirect” occupancy contributes greatly to overall function of enhancers, then predictions of enhancer activity from sequence would be enormously complicated. Even if DNA sequences are in large part predictive of which transcription factors bind regulatory elements, if gene-specific context provides an overriding influence on binding, identification of a transcriptional grammar for these proteins will be fraught with pitfalls.

Relative to this question of how general are the interactions amongst transcription factors, our study clearly shows for the embryonic Dorsal regulon, involving mesoderm and neuroectodermal genes also controlled by Twist and Snail, that the nature of the binding sites, as well as the inferred relationships of cooperativity and quenching (in the case of the short-range repressor Snail) are important elements that determine how particular *cis*-regulatory elements are read out by the transcriptional machinery. Furthermore, our analysis clearly indicates that the list of binding sites in a transcriptional enhancer – is only part of the information to be gleaned. Sophisticated

bioinformatic search approaches have shown that identification of transcription factor motifs is often sufficient to locate novel enhancers, however predicting their functions quantitatively is a much more ambitious goal, and one not likely to be met with simple additive models of transcription factor activity (*22*, *66*, *67*).

Models can be used in different ways to analyze biological systems; in some cases, they can lead to insights regarding the actual quantitative parameters describing reaction rates or molecular interactions, however parameter sloppiness appears to be a common feature of many biological models (*68*). Alternatively, as we demonstrate here, models that differ in details can still be useful in predicting the functional output of a complex system, even as compensated parameters overestimate certain features and underestimate others. We show that the convergence of certain parameter trends in many of the top-rated models does reveal logically consistent insights about how activators and repressors may interact on enhancers. Additionally, and most importantly for extension of these findings to genome interpretation, we see that using groups of models to analyze novel enhancer regions is more effective than a single model. In our case, the differences in performance of parameter sets gleaned from separate fitting attempts were very strongly correlated with the effectiveness of these parameters when predicting novel enhancer regions. This correlation may be weaker in different cases, in which distinct proteins combine according to their own “grammar” to regulate transcription.

The *Drosophila* genome encodes over 700 transcription factors, the majority of which are active in embryogenesis, thus our study has only tackled a narrow slice of the diversity of factors present in this system. We deliberately focused on the early Dorsal

regulon for the richness of quantitative resources available to it, including description of levels, binding specificity of trans-acting factors, and genomic data on *in vivo* targets. Other groups of transcriptional activators are not currently as well curated as this, but with continuing advances in genomics and high-throughput technologies, we foresee the development of relatively complete descriptions of gene and protein expression, as well as genomic protein occupancy and genomic variation for a majority of the most important transcription systems in the embryo. Similar advances are clearly developing these panoramic views in vertebrates as well, and it is possible that part of the "personalized genome" efforts will provide a basis for similar modeling in this domain in humans. As successful as this modeling effort was to understand one regulon, several areas require further work, however. First, we focused on genes activated by Dorsal in the mesoderm and neuroectoderm, but did not extend this to a special class of enhancers in which Dorsal binds in a special context with the Cut and Dri proteins that permits it to recruit Groucho cofactor and function as a repressor. These elements are active only on dorsal ectodermal genes. The focus on ventrally and laterally active promoters allowed us to maintain Dorsal in the role of an activator. Second, many transcription factors mediate both repression and activation, depending on cellular signaling pathways. Such context dependence indicates that one must have information about the "state" of signaling systems before a proper interpretation of the DNA content of an enhancer can be performed.

The nature of models used in thermodynamic applications has been insufficiently investigated in past studies. Previous thermodynamic applications to complex Drosophila developmental modules were sufficient to demonstrate that binding site

information can be incorporated in such approaches to display real or simulated gene expression. These efforts were limited by a complete absence of distance-dependence in interactions, in one case, or a limited suite of possible interactions in another (*21*, *22*, *33*). Previous thermodynamic models such as those employed by Segal et al. used rather simplified assumptions for protein-protein interactions, limiting cooperative interactions to homotypic types. Our modeling approach breaks new ground in a number of levels. The specific antagonistic contributions of the short-range Snail repressor is represented in our models as a direct reduction of activator potential, rather than an independent negative input to the basal machinery, which does not correspond with know biological processes. Although we explicitly consider the impact of knowledge of transcription factor binding site preferences (Fig. 3.6), testing model outcomes when using the slightly different PWMs derived from distinct types of experiments, we do not attempt to improve the fit by altering the predicted binding specificity of the factors, which was a biologically unsupported approach used by Segal *et al.* (*22*). We expect that biophysical properties of these proteins will be similar when interacting with DNA segments throughout the genome. Our models also include specific terms for heterotypic cooperative interactions, in this case between Dorsal and Twist. It is interesting that these parameters appear to be among the less sensitive of all estimated parameters, indicating that models are not strongly influenced by changes in values of these parameters. A major difference in the predictions between our models and that of Segal et al. is that they conclude that weak interactions, from sites with near background affinity levels, contribute significantly to overall expression. This concept is difficult to test experimentally, as mutation of near-background sites is not expected to

change output significantly, either because there is little binding in the first place, or that binding to the mutant sites will be sufficiently similar to the original sites that expression will not be strongly impacted. We do test their model of weak site activity by predicting expression of the 38 rhomboid constructs fit by 120 models when using lower threshold for Dorsal, Twist and Snail binding. In these cases, the overall fit deteriorated (Supp. fig. S3.6).

By implementing a variety of schemes describing cooperative and quenching functions, we cast a wide net that might allow us to recover parameters that actually provide information on biochemical activities of the proteins we study here. We note that some of the recurring trends observed for many of the most successful models are likely to reflect biochemical interactions occurring on enhancers. First among these is the long-range cooperative interactions for Dorsal predicted by a number of models; our initial observations that individual Dorsal binding sites contribute similarly to overall output may be consistent with such interactions. Long-range interactions may be consistent with the general, relatively 'distance-independent' cooperativity noted earlier (*57*). The biochemical mechanism underlying such cooperativity may involve antagonistic binding with nucleosomes (*69*). Such cooperativity is almost never predicted for the Snail repressor, which like Dorsal has four binding sites in the enhancer; the strongly differential contribution of individual Snail sites noted in our study would indicate that these proteins have more of an additive effect on repression. Rather than a generic repression contribution, as assumed in previous studies, however, the positioning of Snail sites would have paramount influence on how each binding event impacts enhancer function (*22*). We also found a common theme of monotonically

decreasing distance-dependence for quenching in many (but not the majority of) parameter sets. The generally low values for Snail cooperativity that we frequently recovered suggest that the contributions to gene expression of individual enhancer states containing multiply-bound Snail repressors are underrepresented. Thus, binding by an individual Snail protein is predicted to be more common, and important in dictating repression. This notion is supported by chromatin studies. The Knirps repressor, which interacts with the same CtBP corepressor that binds to Snail, is observed to compact chromatin locally and may interfere with multiple binding events of this repressor, in contrast to the cooperative effects that likely aid multiple activator occupancy (*70*).

The high degree of parameter compensation possible for many of our models suggests that alternative descriptions of quenching by distance may be balanced out by differing Snail scaling factors. It is striking, however, that using PWM settings that accurately reflect footprinted Snail sites is a major determinant of model performance, therefore, there must be a limit to the extent of such compensation (Fig. 3.6). Our analysis was not set up to identify certain specific features that pertain to Dorsal and Twist action on certain enhancers. For instance, the orientation of a Twist site closely juxtaposed to a Dorsal site in the *vein* enhancer was demonstrated to have a strong influence on activity (*33*). The general importance of such orientation is unknown; our models do not consider this feature, thus to the extent that they succeed, this effect may only impact some *cis* elements. The tight coupling of this same Twist-Dorsal site was also noted to strongly impact readout of the enhancer, thus this pair of motifs might

represent a highly sensitive, enhanceosome-like element within the *vn* enhancer (*55*). The generality of this pairwise interaction is unknown.

To simplify modeling, our study focused on three of the most important transcription factors that are necessary for tissue-specific quantitative expression of the embryonic Dorsal regulon. Nonetheless, additional factors may impact the activity of some or all of these enhancers. First, recent studies identified the Zelda transcription factor as an early-acting transcription factor that binds to many enhancers of the blastoderm embryo, and whose activity is critical for wild-type activity of these elements. Genes such as *sog*, and to a lesser extent *rho*, show delayed and weakened expression in embryos lacking maternal Zelda (*63*). We focus on a time point at which many elements are transitioning from Zelda-dependent to –independent expression, but in general one might expect that our models overestimate the activation potential of Dorsal and Twist to compensate for this missing input. In fact, we note that the enhancers that are predicted to feature highest levels of Zelda do show greater divergences in predicted expression, possibly because of this effect. Second, we note that the spatial regulation of some enhancers is critically dependent on additional input. The *sim* midline enhancer, which shares with *rho* the overall strategy of Snail repression and activation by Dorsal for mesectodermal expression is known to respond to additional inputs from the Notch signaling pathway via Su(H) sites to refine its single-cell-wide expression pattern (*71*, *72*). Our predicted output of this enhancer is consistently too wide, as expected.

A third likely source of divergence for our models is the possible impact of chromatin structure on the availability of DNA binding sites. Although there is

considerable controversy about the extent to which DNA sequence influences nucleosome positioning, the arrangement of nucleosomes in transcribed regions can be strongly influenced by overall nucleotide sequence, which in some cases appears to be sufficient to prevent high-affinity activator sites from driving gene transcription (*73*, *74*). Implementation of nucleosome occupancy in modeling approaches such as ours can be based on the average nucleotide composition of the enhancers in question. The predicted nucleosome occupancy might be used to revise the score of particular regions so that binding sites would have greater or lesser potential compared to nucleosome-free regions. Nucleosome occupancy can also be influenced by the action of transcriptional activators and repressors (*75*), but in this case we are likely to detect these interactions through the values of cooperativity or quenching parameters that reflect both direct and indirect interactions between transcription factors.

Acknowledgments

We acknowledge Michigan State University High Performance Computing Center and the Institute for Cyber Enabled Research for assistance with computational analysis, Dr. Melinda Frame (MSU Center for Advanced Microscopy) for assistance with confocal microscopy and Nicholas Panchy for help with sequence analysis, Anne Sonnenschein, Rewatee Gokhale, Max Winkler and Ramona Beckman for assistance with cloning of constructs, Dr. Robert Zinzen and Dr. Dmitri Papatsenko for help with DVEx dataset, Arnosti Lab members for helpful discussions, Dr. Yuehua Cui for advice on statistical analysis, Martin Scherr for assistance with Mathematica coding, and Dr. Chichia Chiu and Dr. Ian Dworkin for comments on the manuscript.

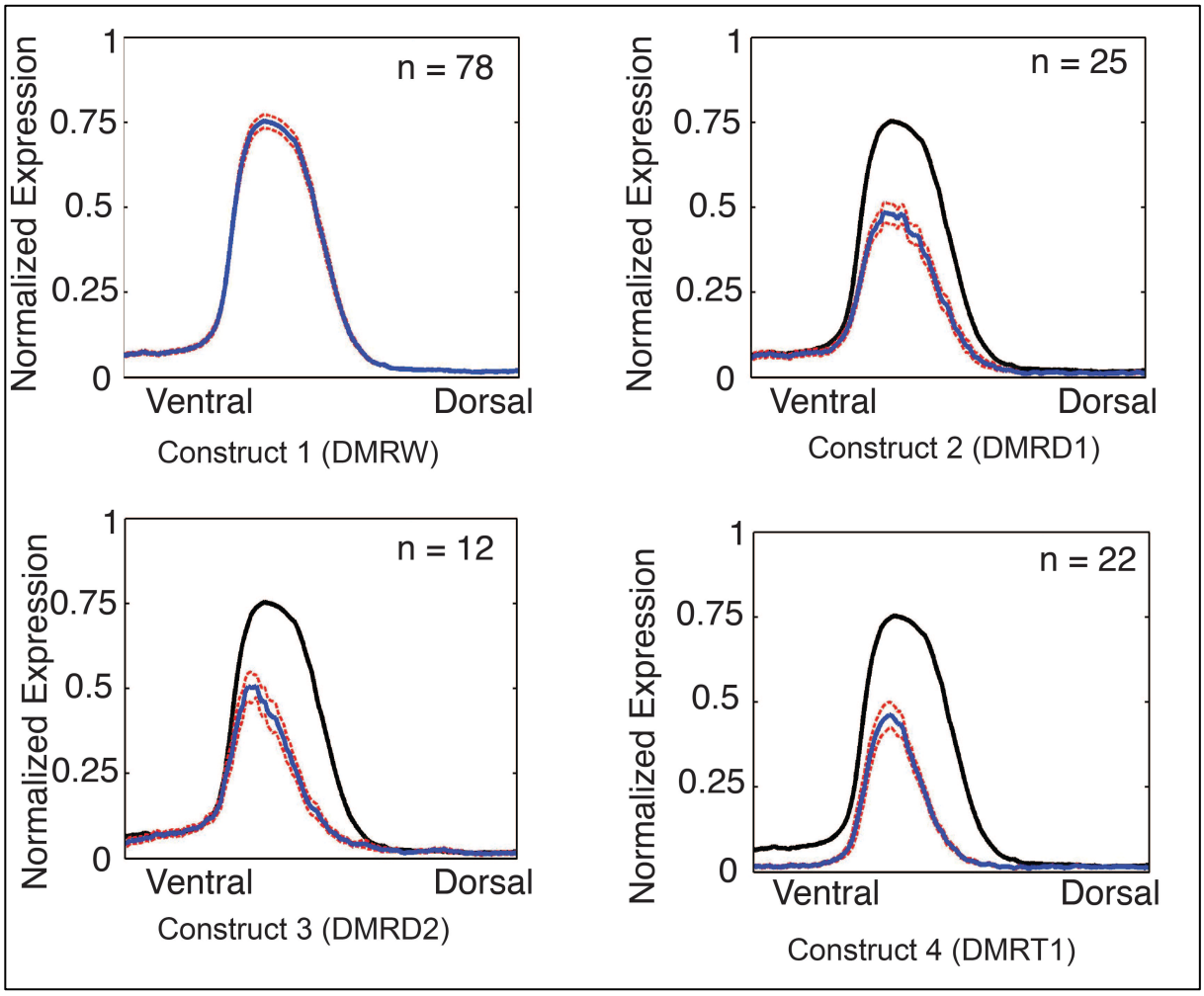


Figure S3.1 Average expression plots of enhancer constructs. Each panel shows the averaged plot (blue line) for the construct. The red dotted lines show the spread of standard error for expression. Box on upper right shows the number of embryos averaged for each construct. See table T1 for details on nomenclature of constructs.

Figure S3.1 cont'd

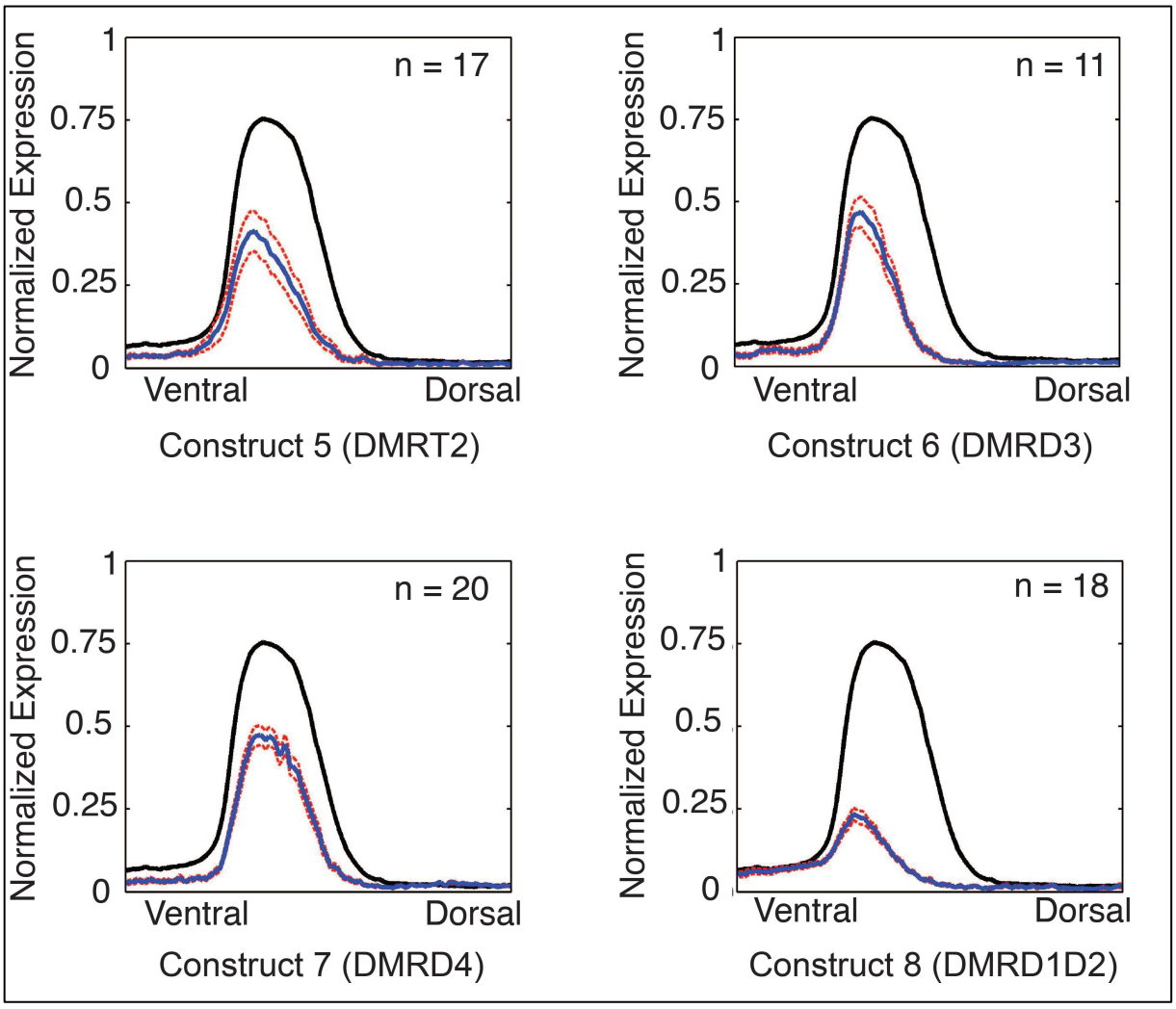


Figure S3.1 cont'd

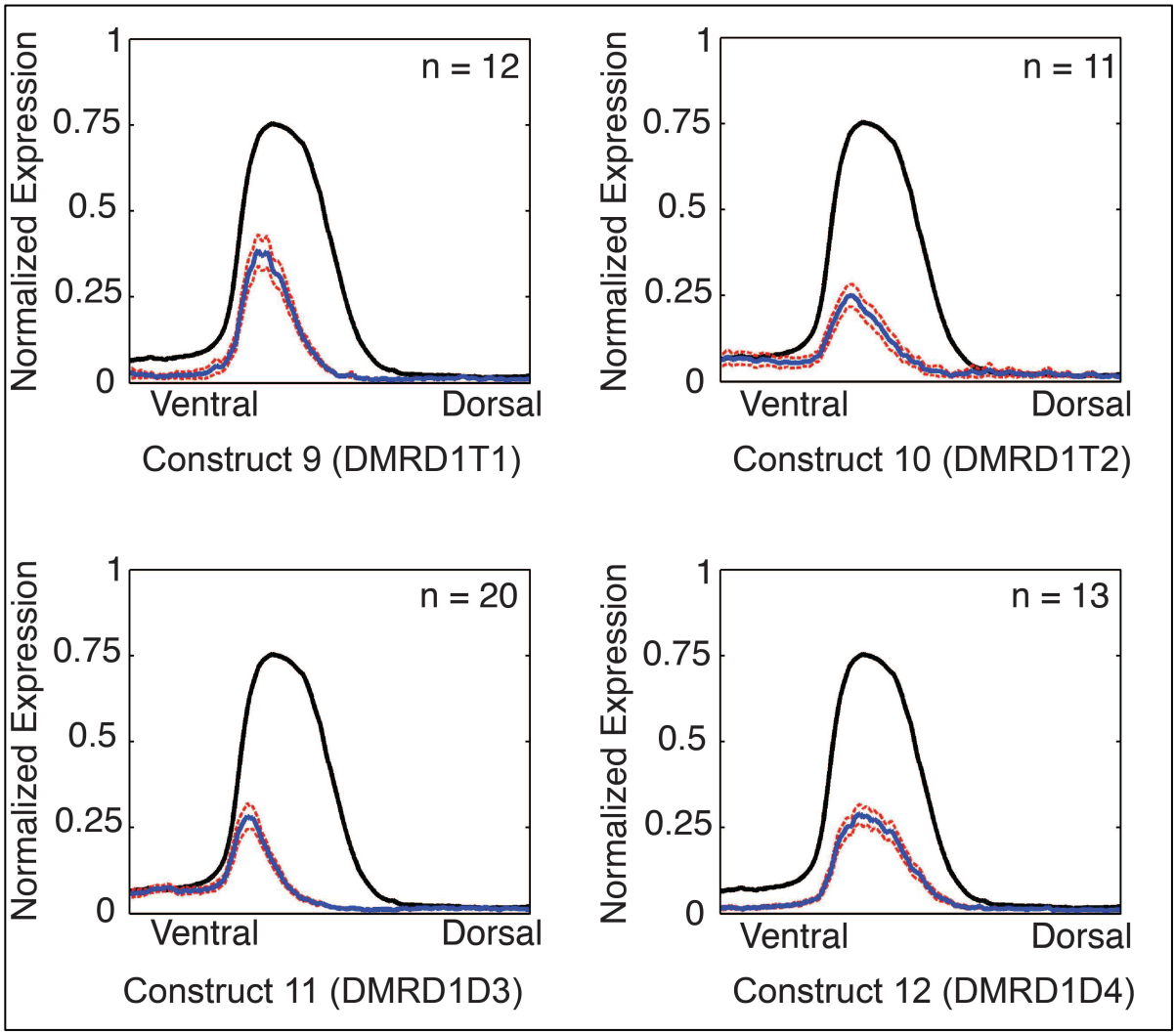


Figure S3.1 cont'd

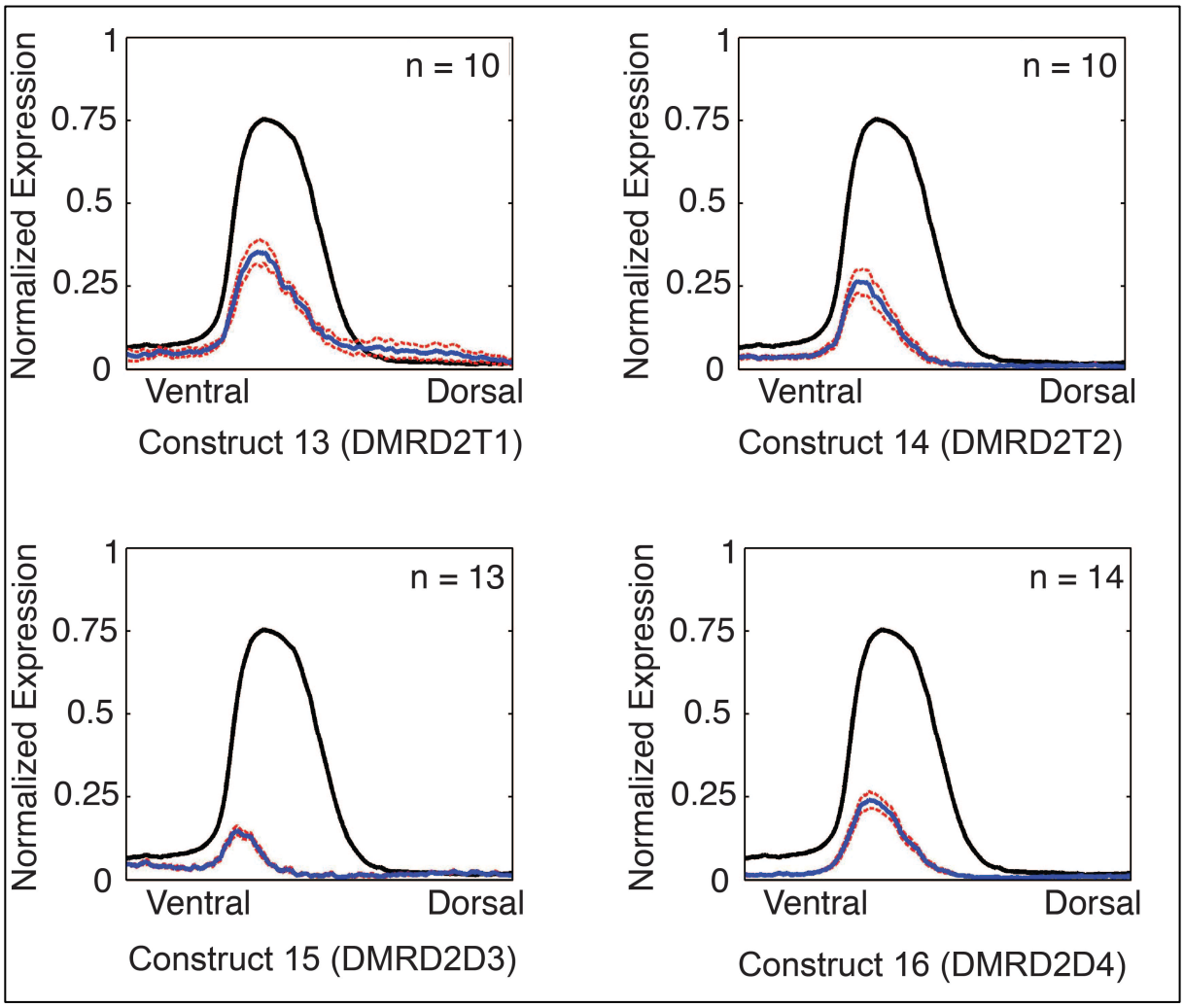


Figure S3.1 cont'd

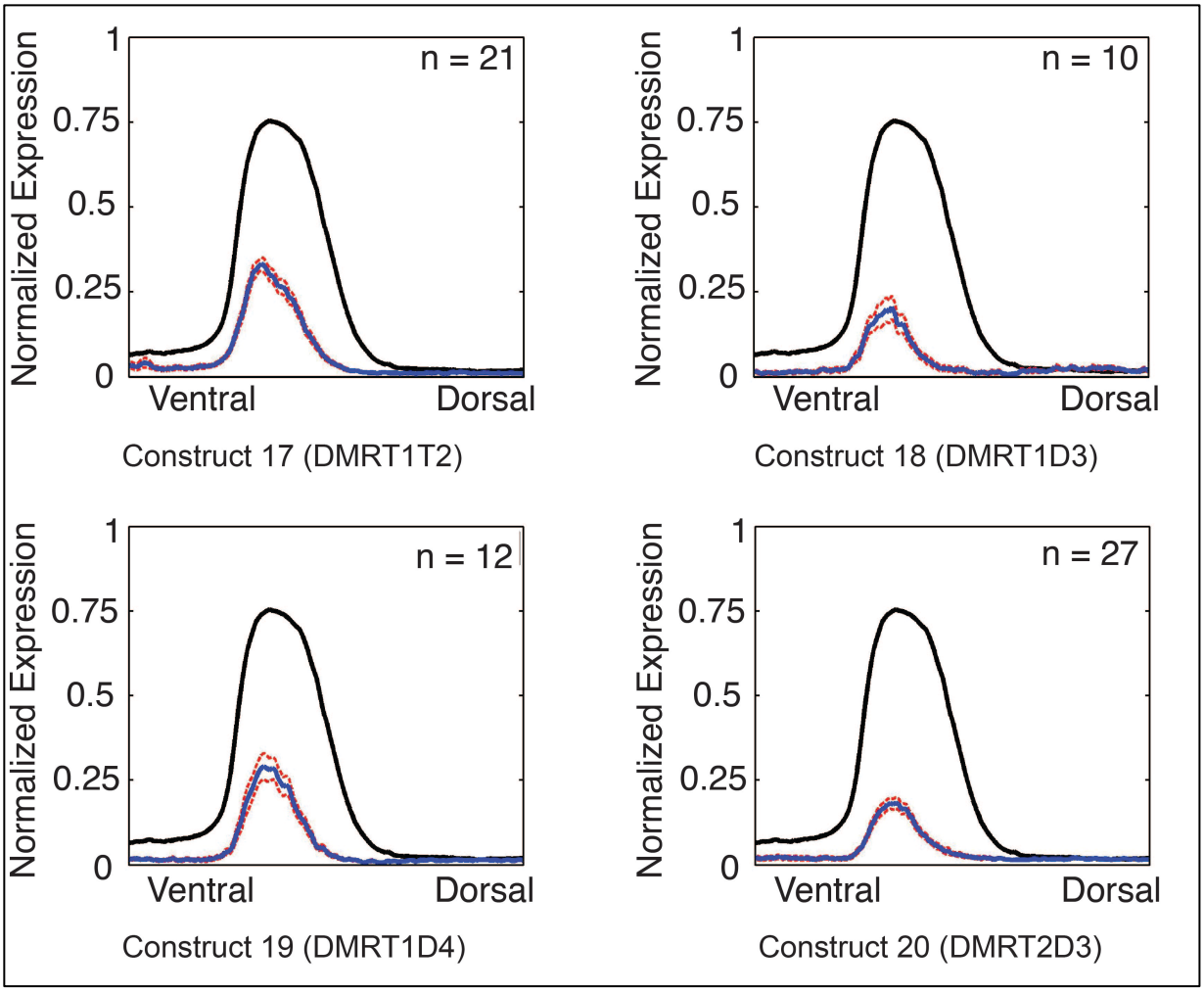


Figure S3.1 cont'd

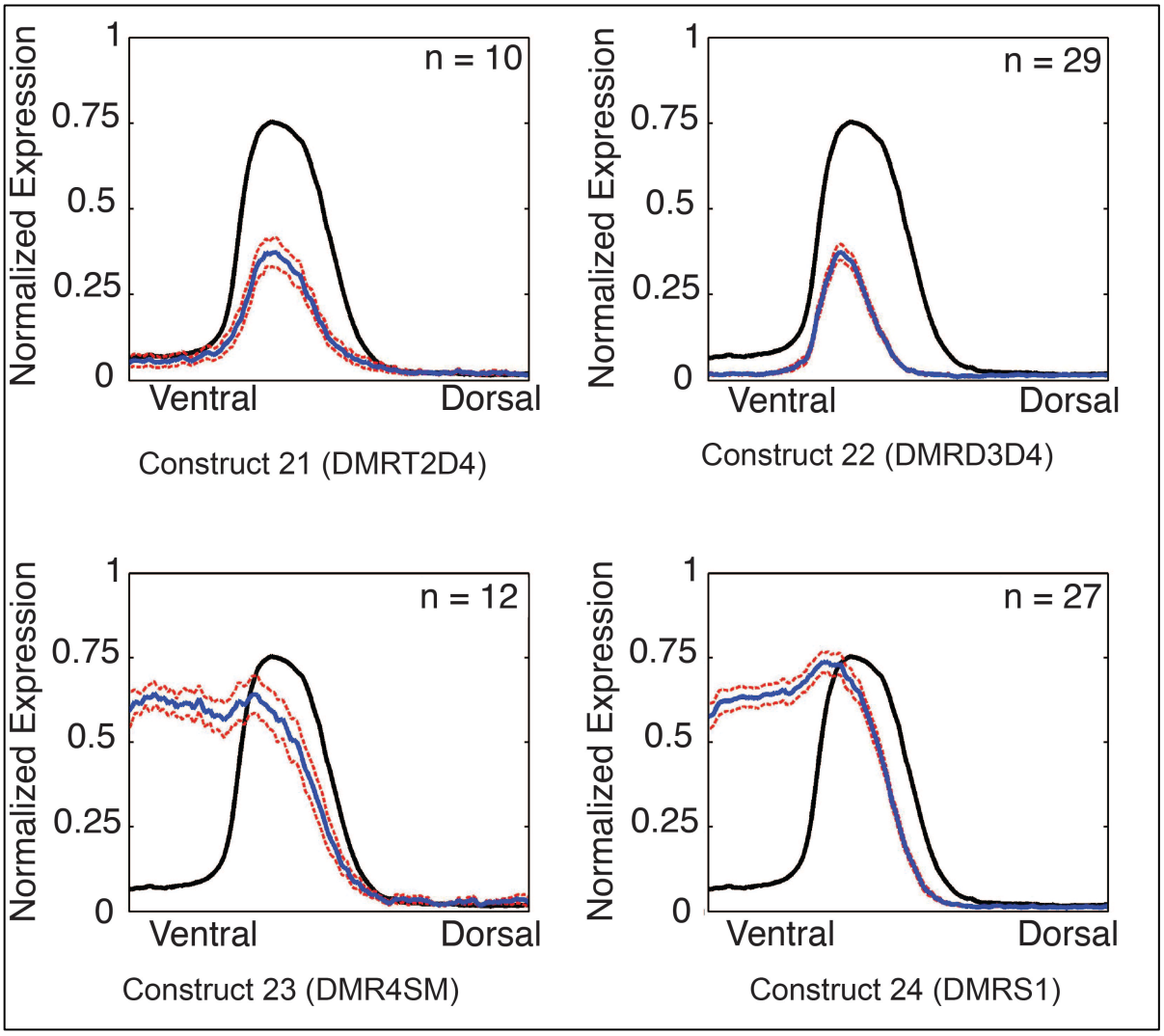


Figure S3.1 cont'd

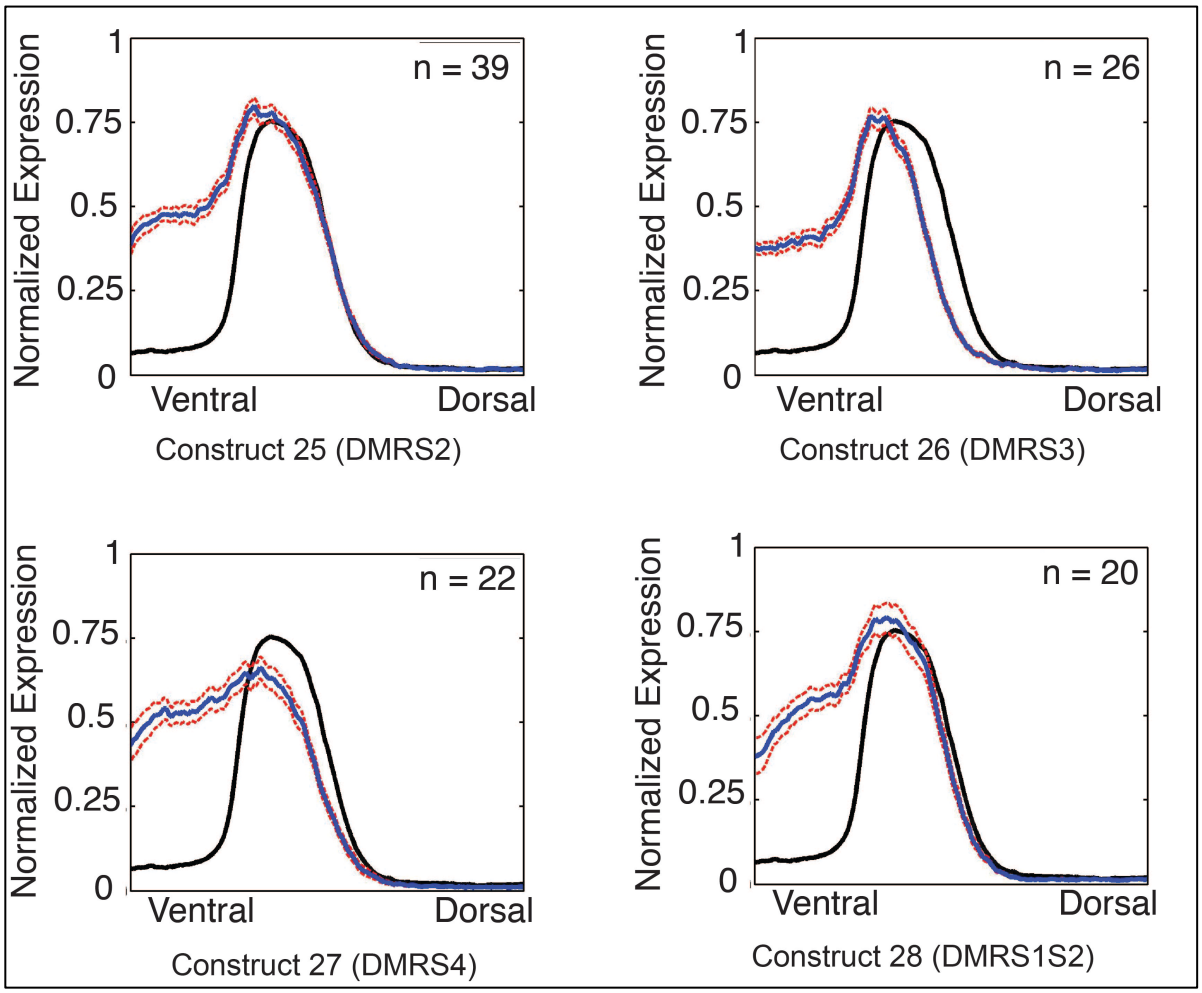


Figure S3.1 cont'd

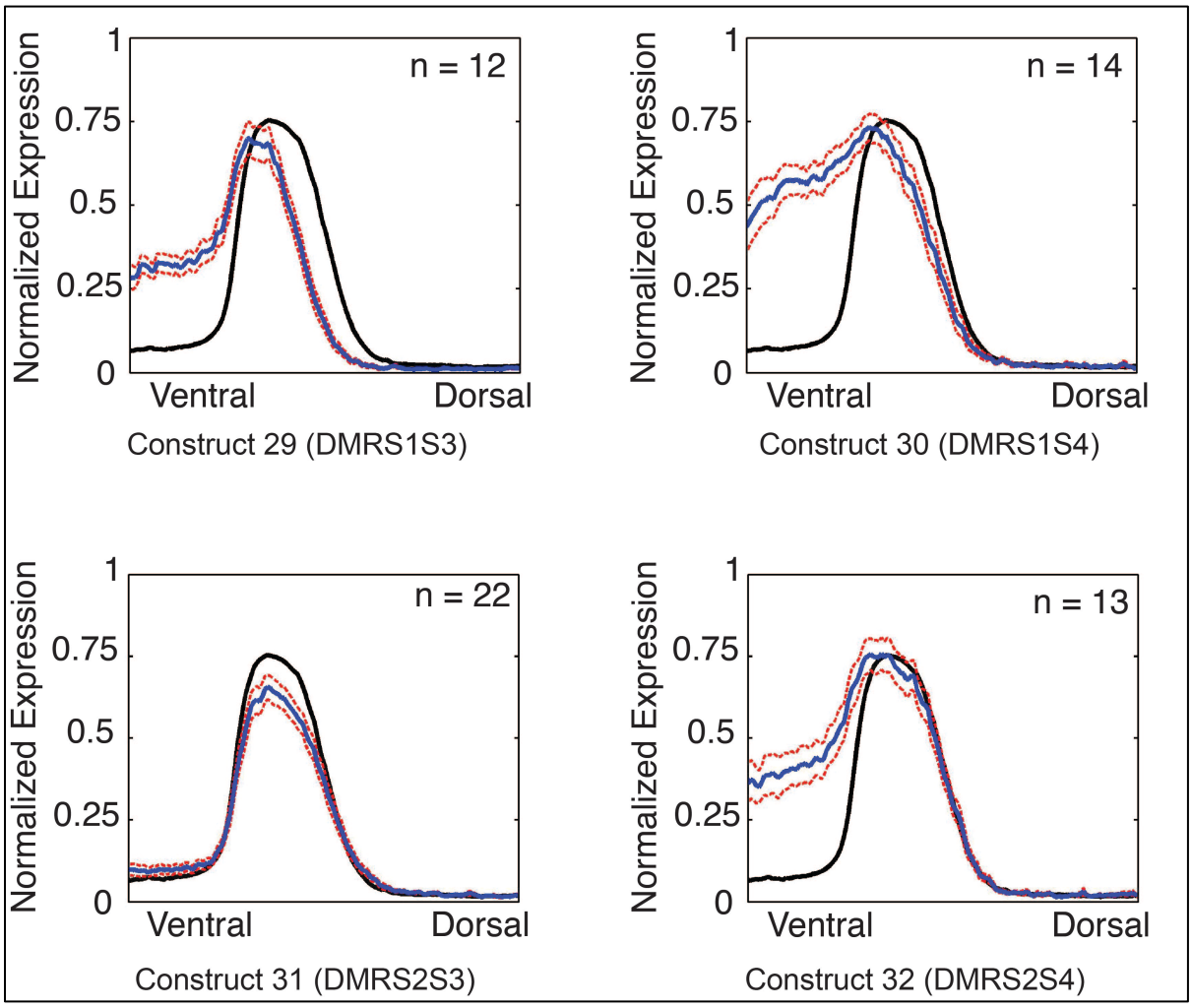


Figure S3.1 cont'd

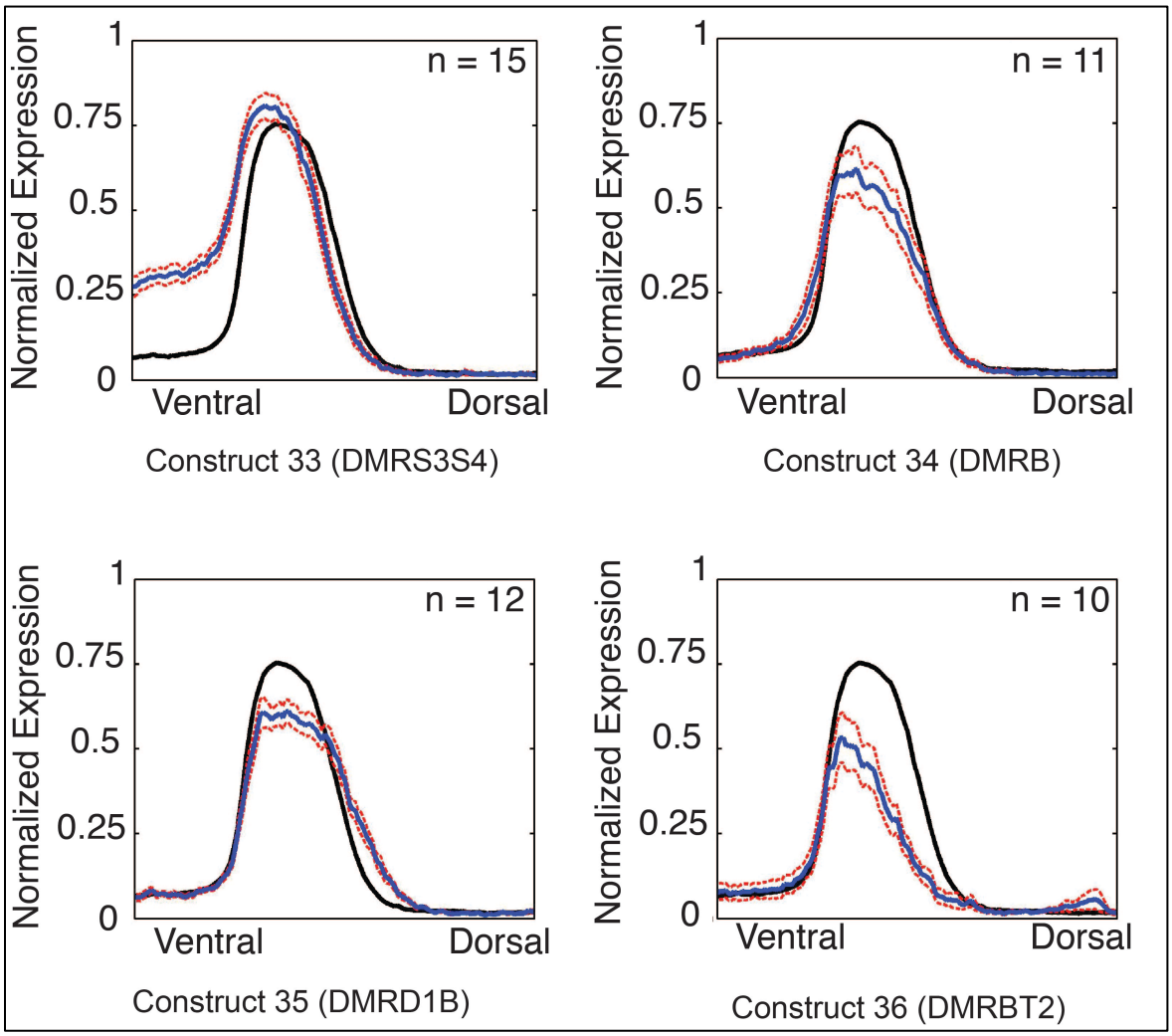


Figure S3.1 cont'd

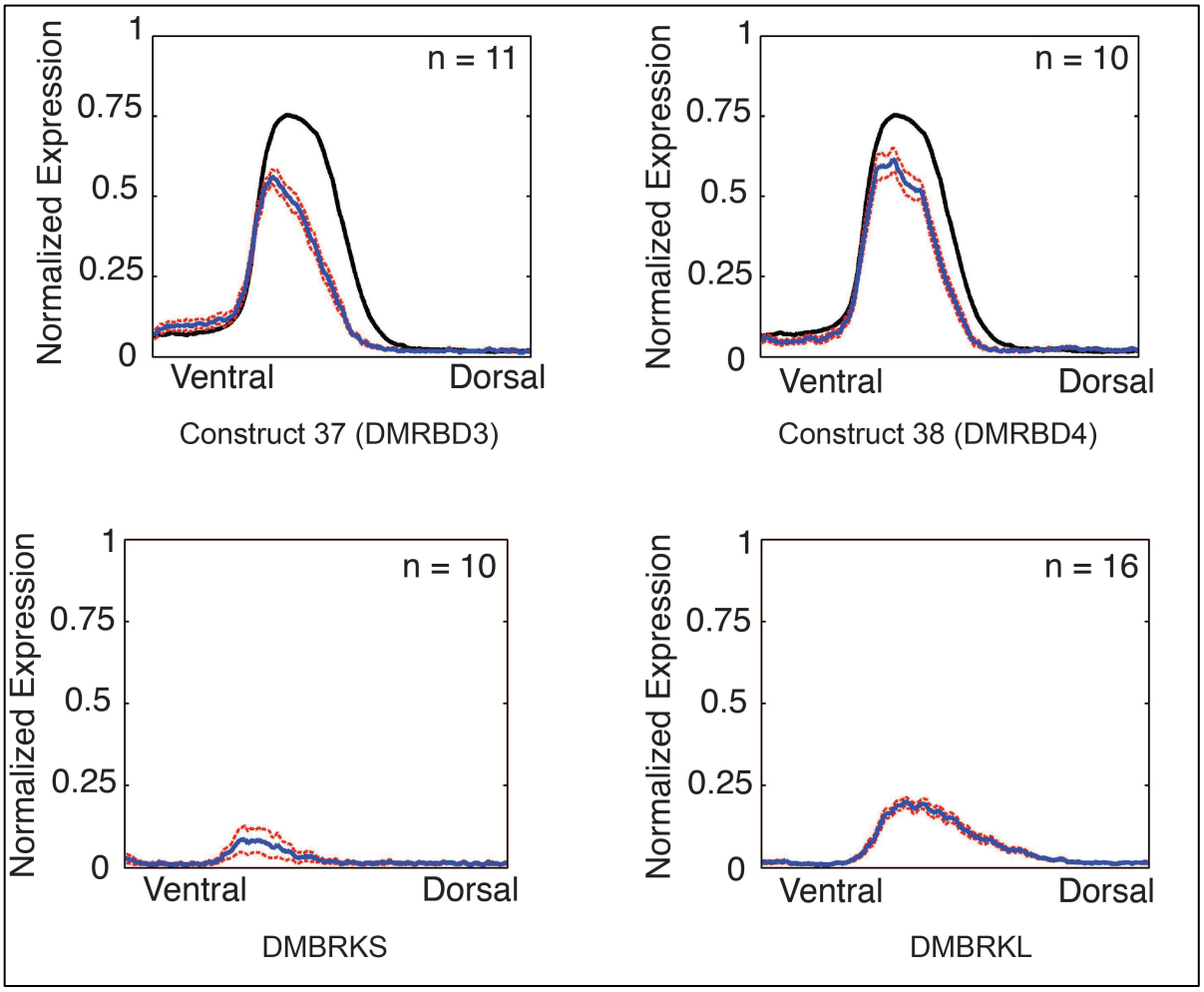


Figure S3.1 cont'd

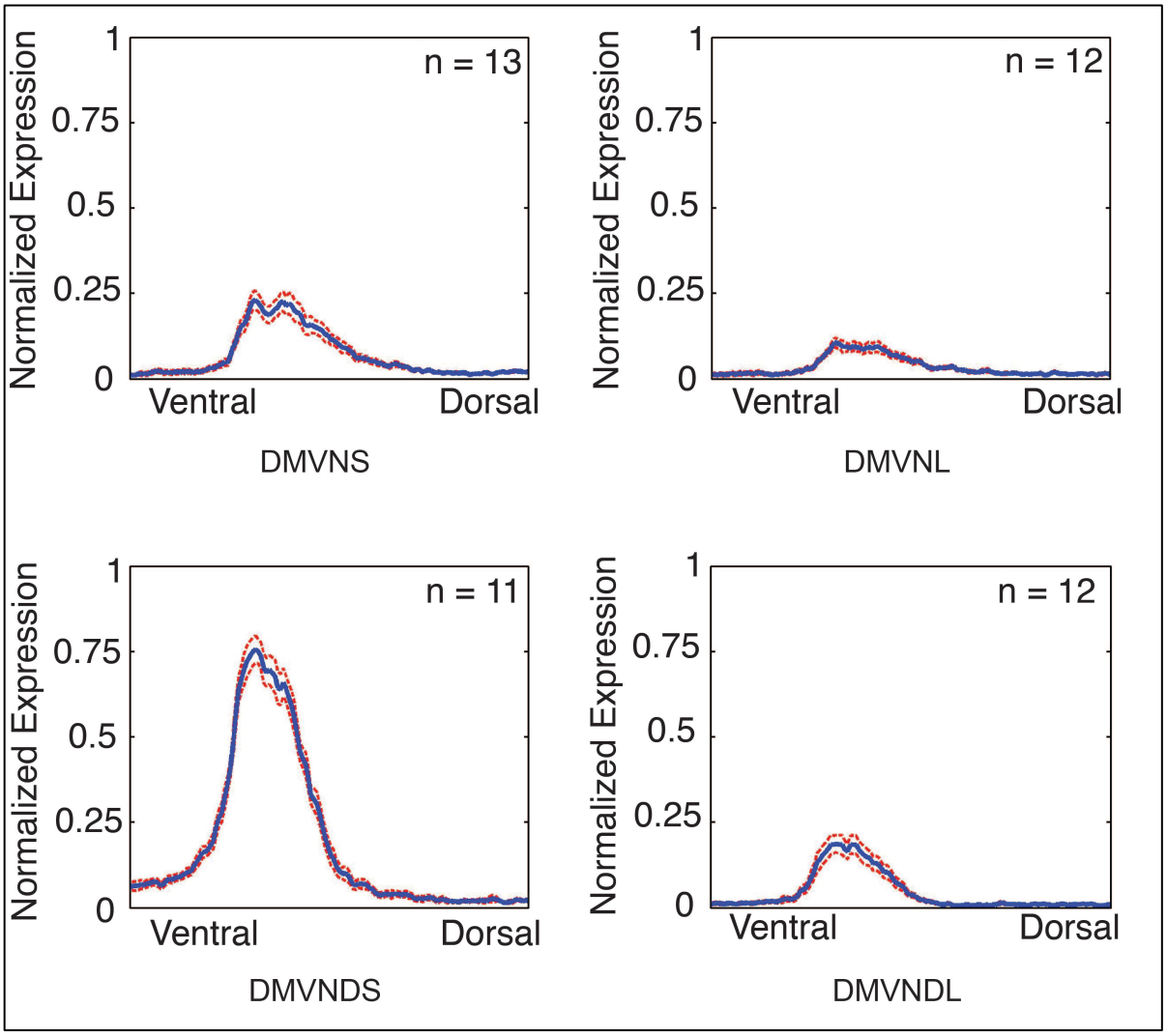


Figure S3.1 cont'd

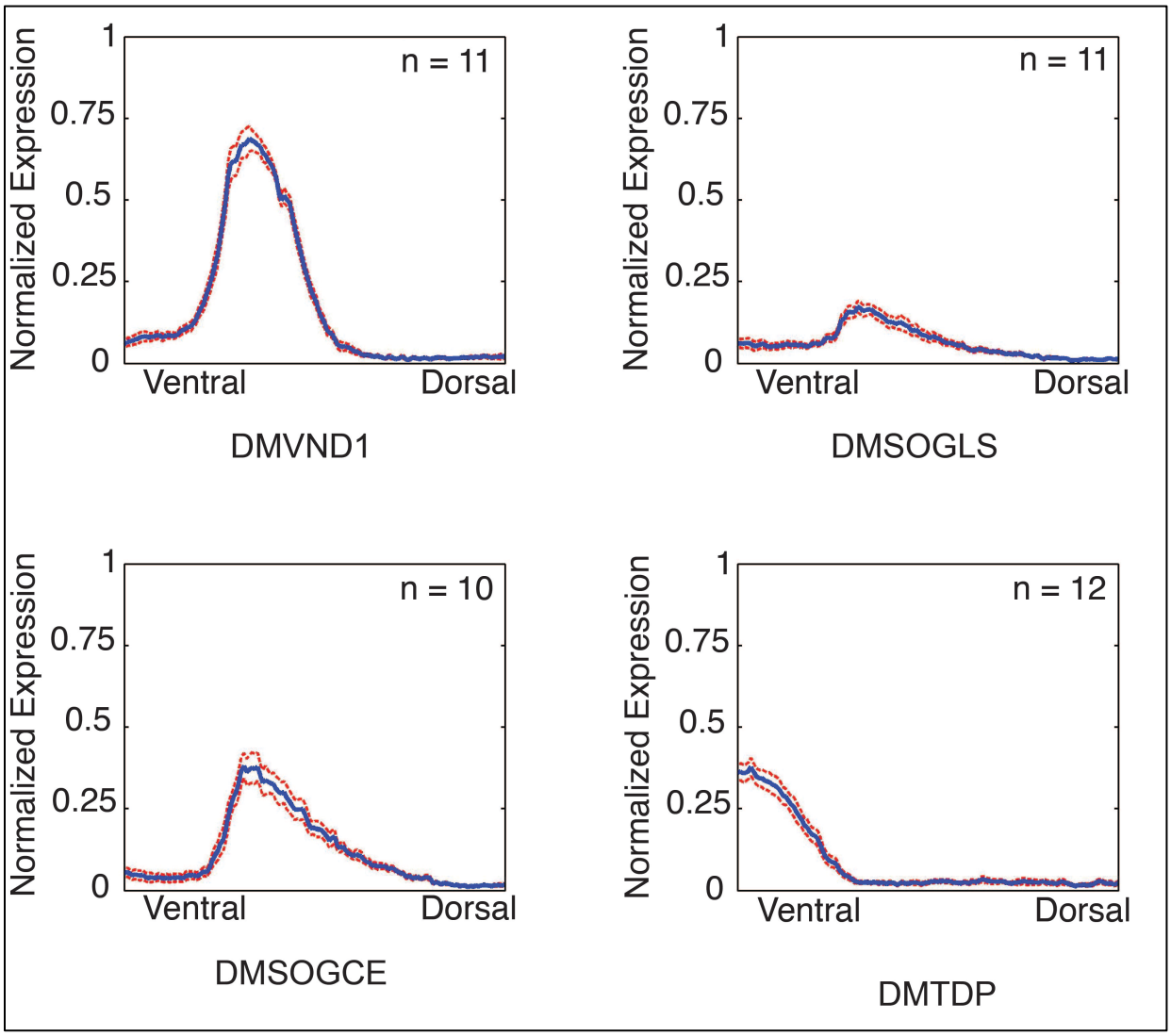


Figure S3.1 cont'd

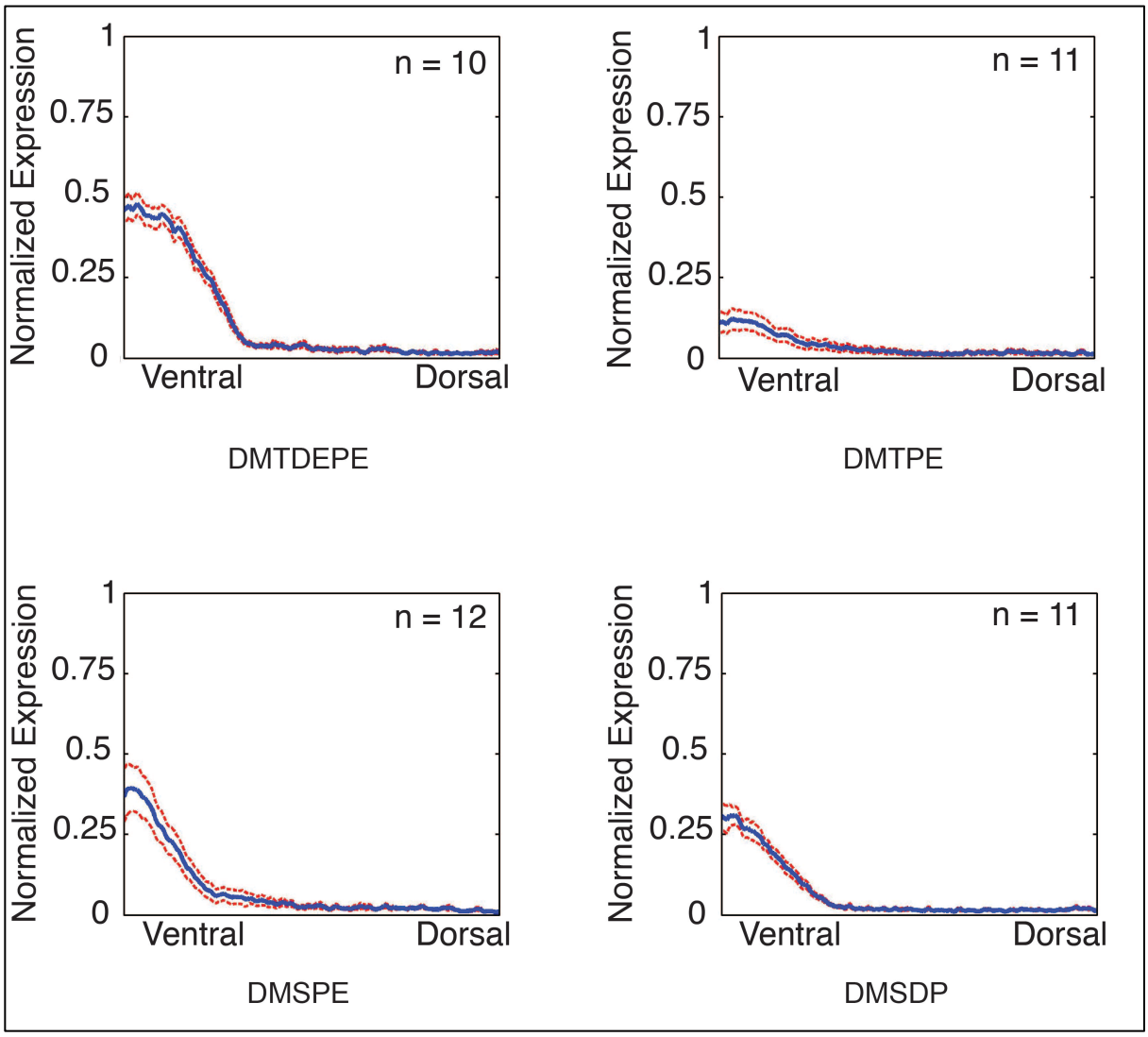


Figure S3.1 cont'd

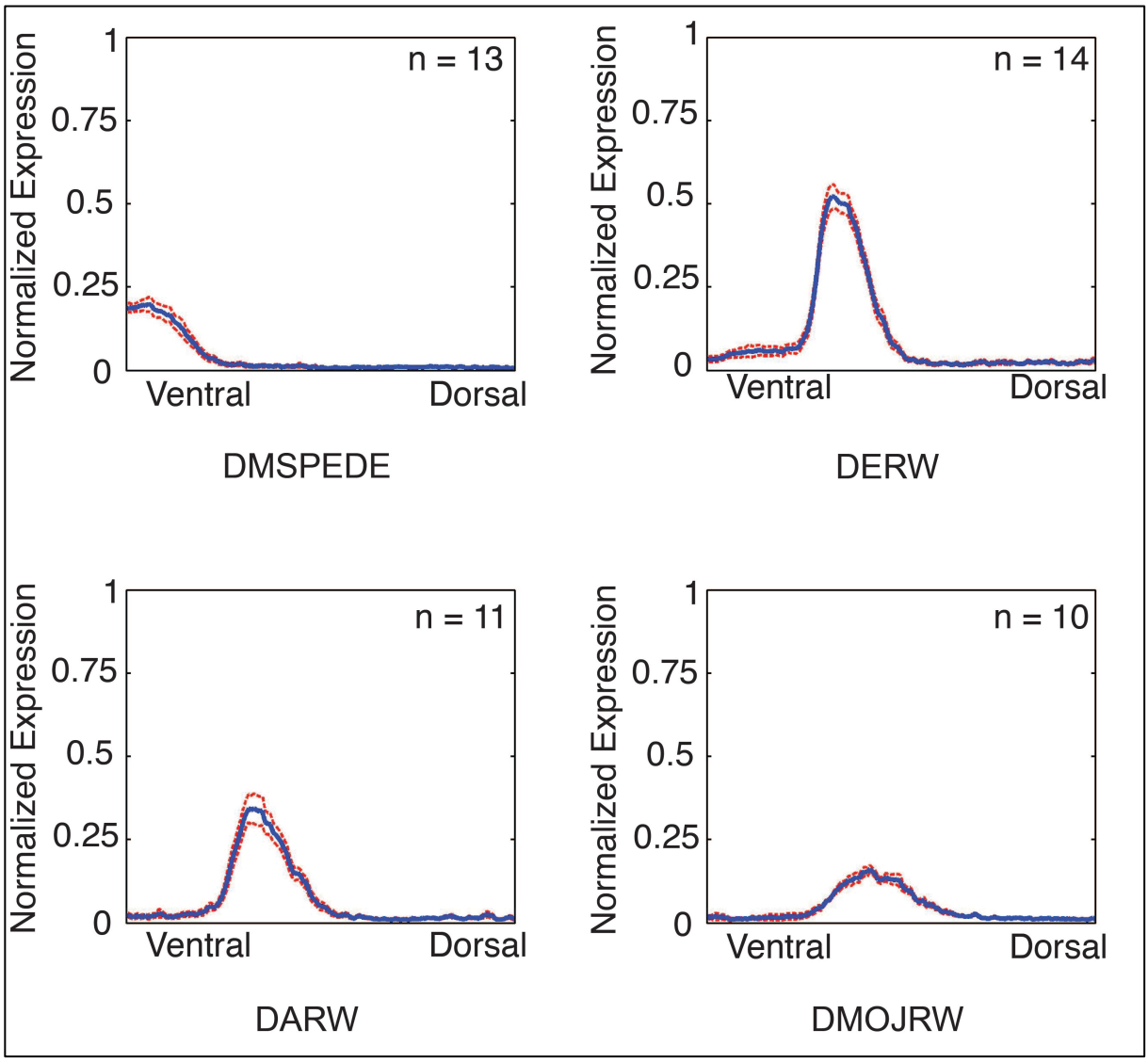
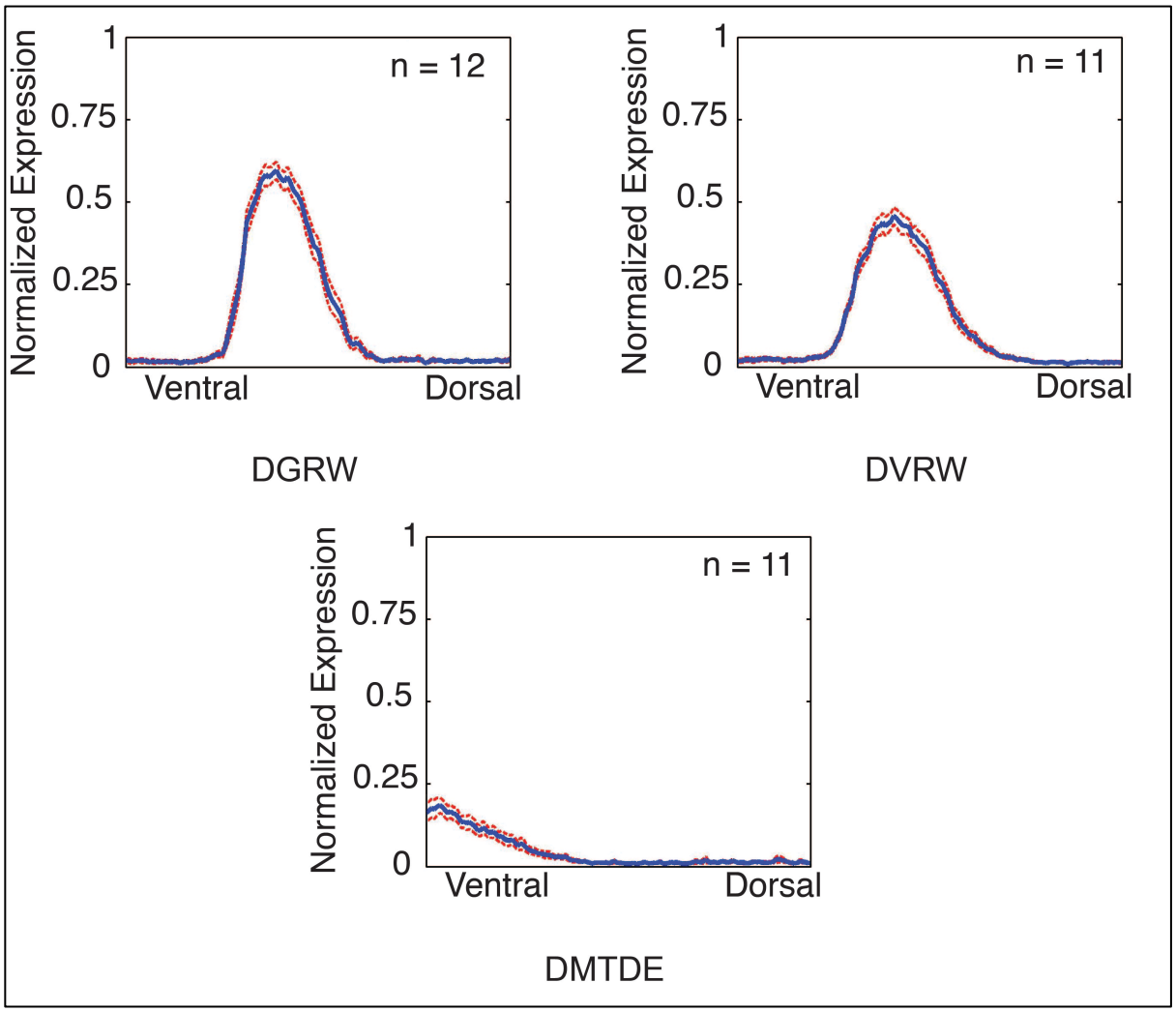


Figure S3.1 cont'd



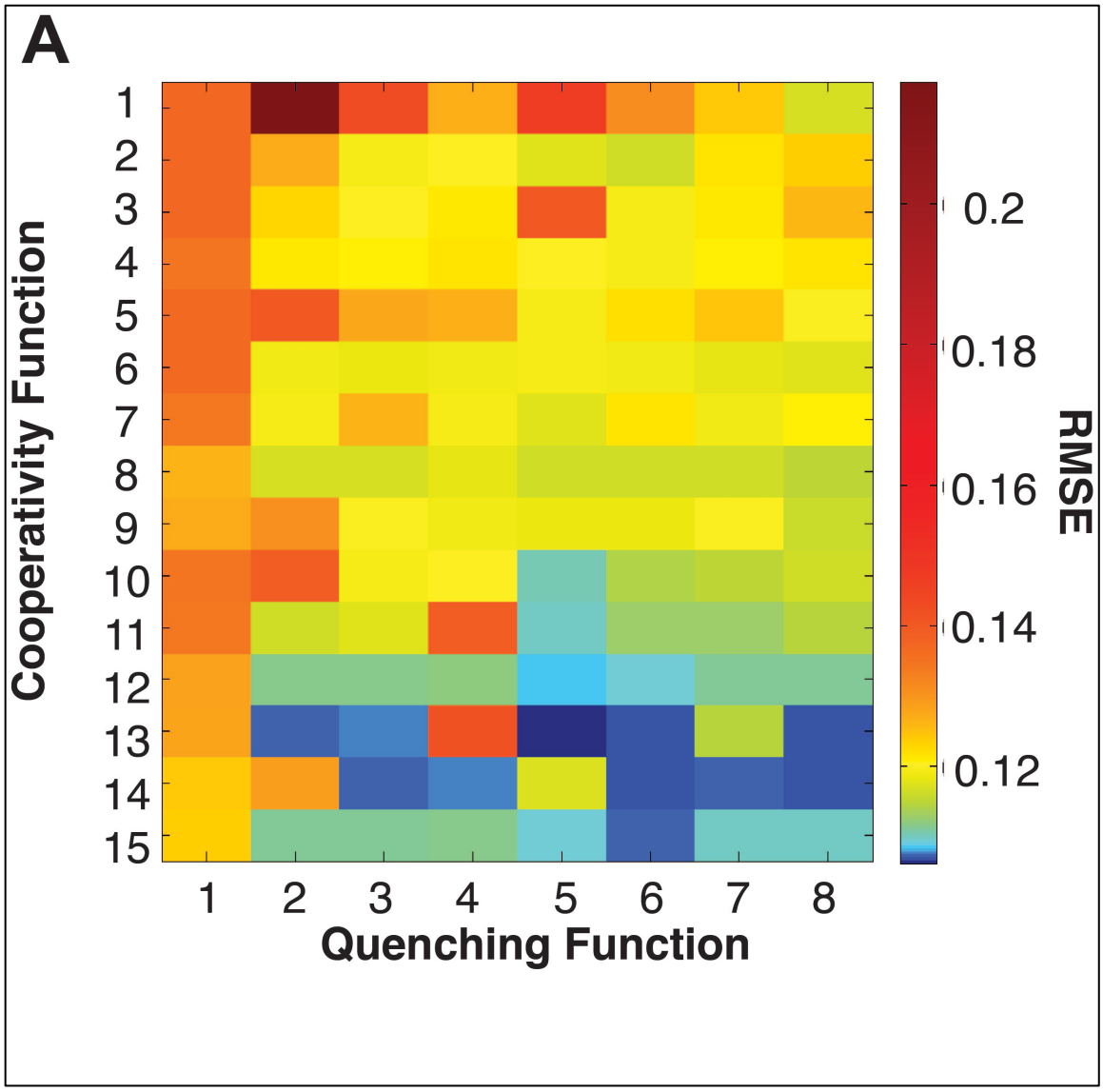
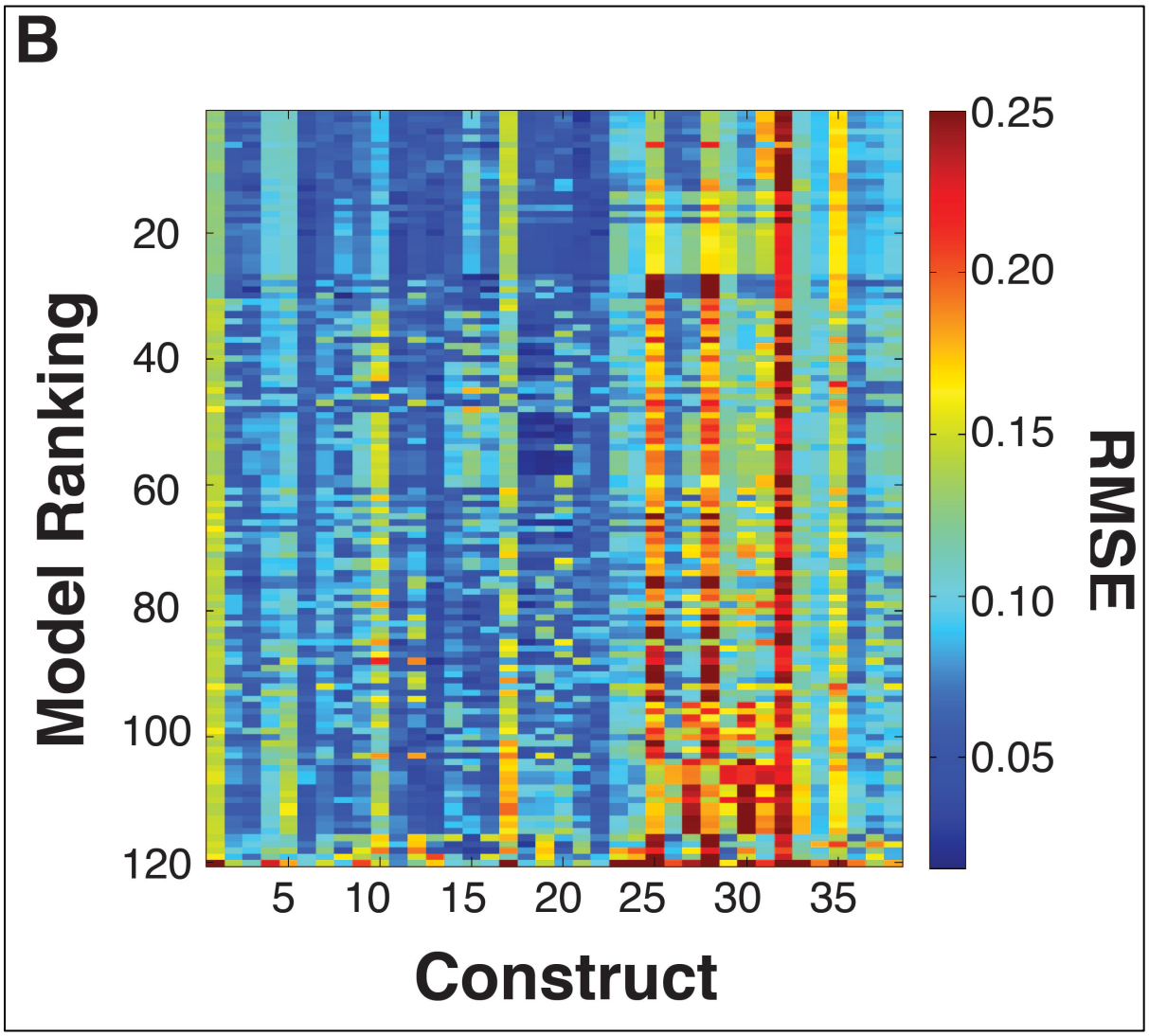


Figure S3.2 Global and construct-wise performance of 120 models on a different PWM setting. (A) Heatmap shows performance of 15 cooperativity models (Y-axis) and 8 quenching models (X-axis) on 38 enhancer constructs. Performance of models is indicated by color-coded RMSE values (bar on right).

Figure S3.2 cont'd



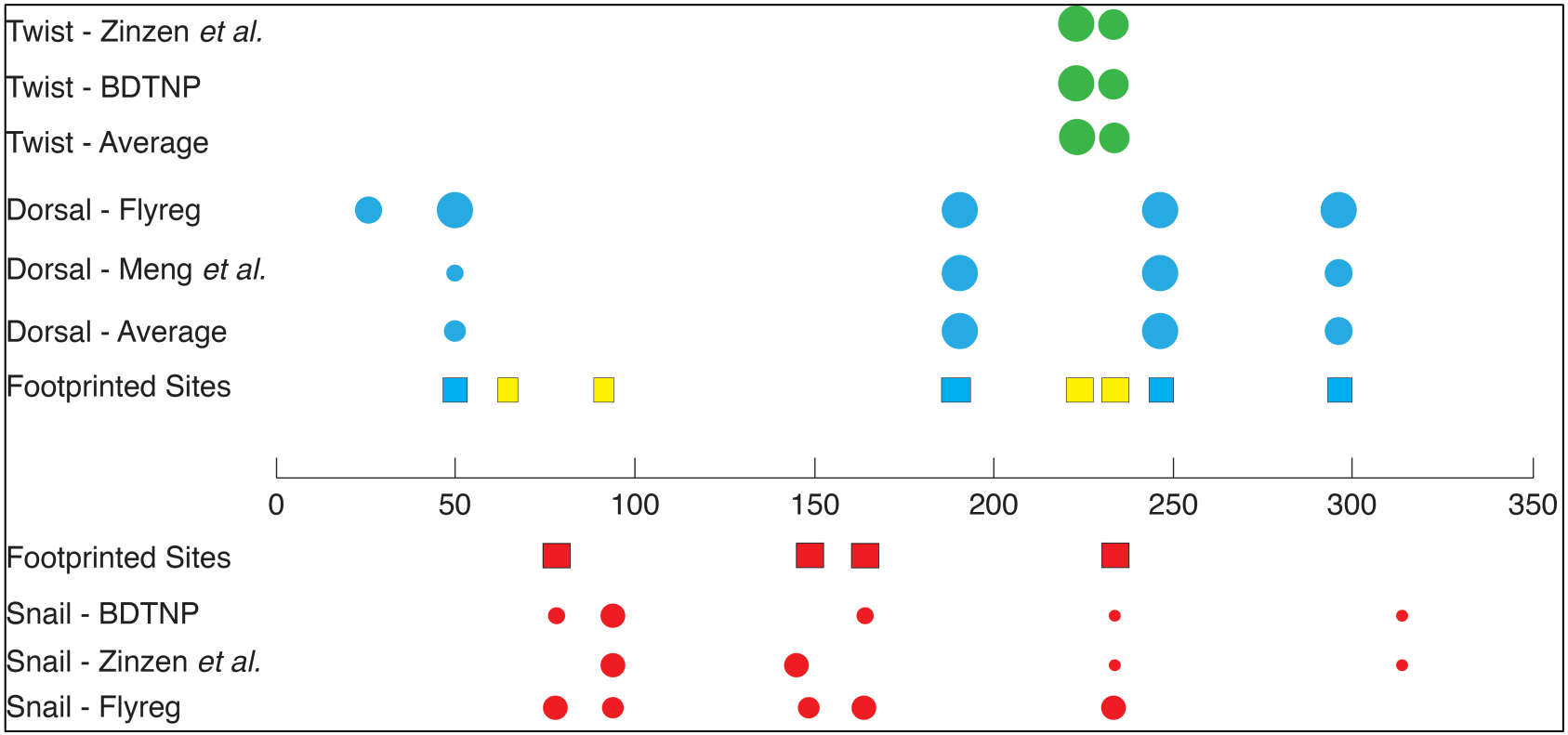


Figure S3.3 Landscape of binding sites on *rho*NEE enhancer. Boxes show different footprinted binding sites for Dorsal (blue), Twist/bHLH (yellow) and Snail (red) on the enhancer. Circles show predicted binding sites using different PWMs for Dorsal (blue), Twist (green) and Snail (red). The sizes of circles correspond to MAST scores of predicted binding sites.

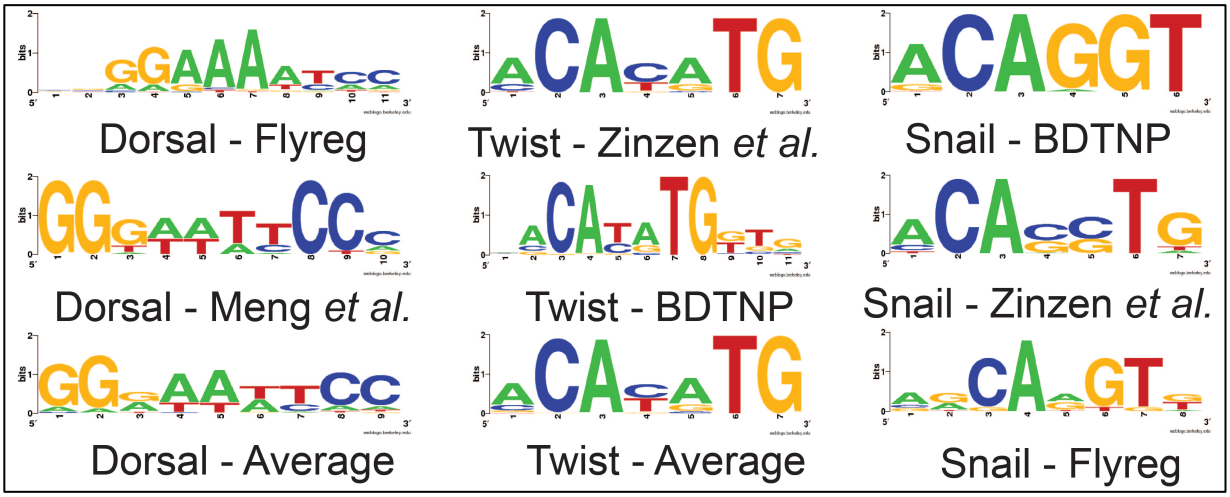


Figure S3.4 Position weight matrices derived from different sources. Panels show different PWMs used for Dorsal (left), Twist (center) and Snail (right) to score binding sites.

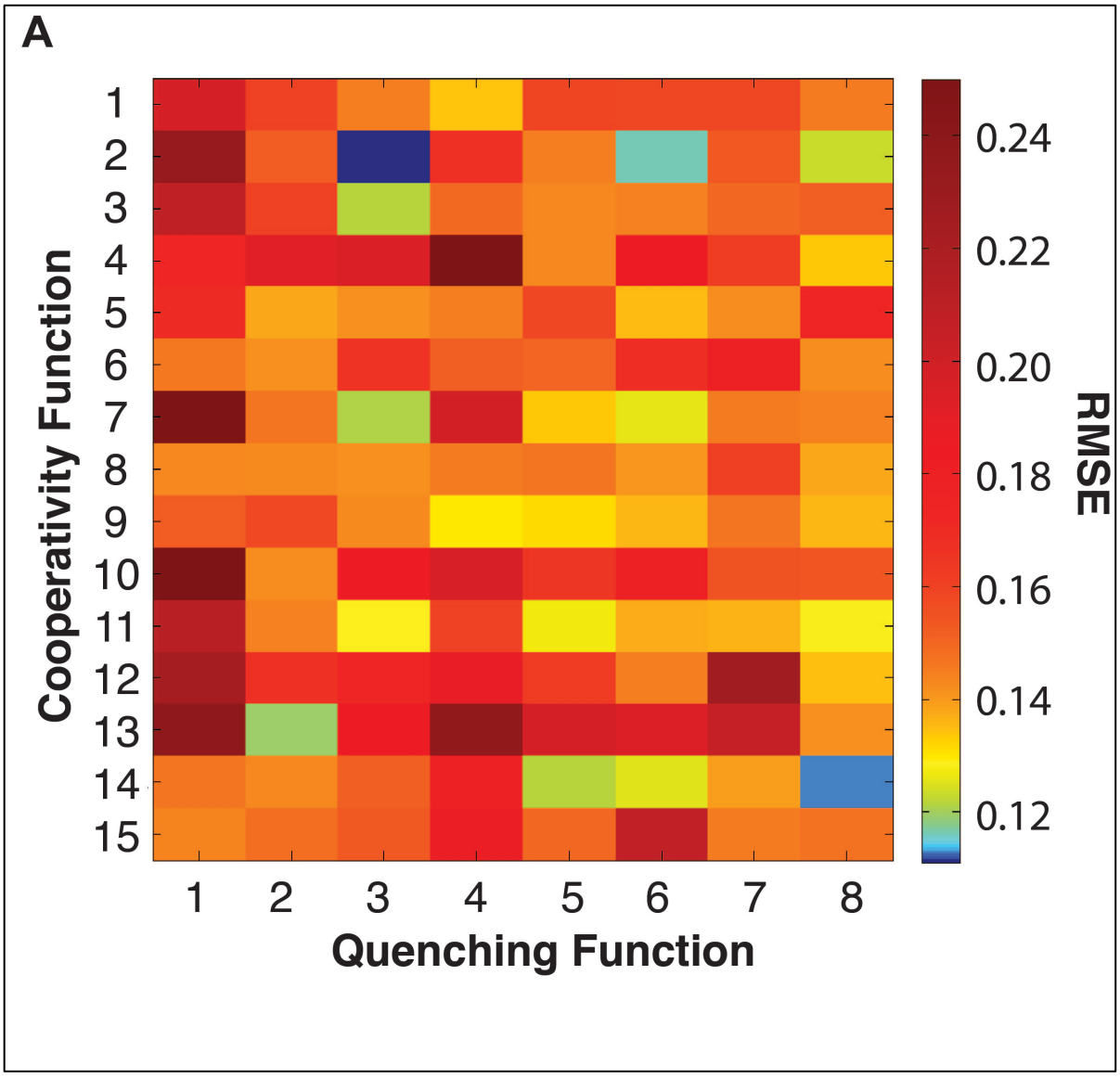
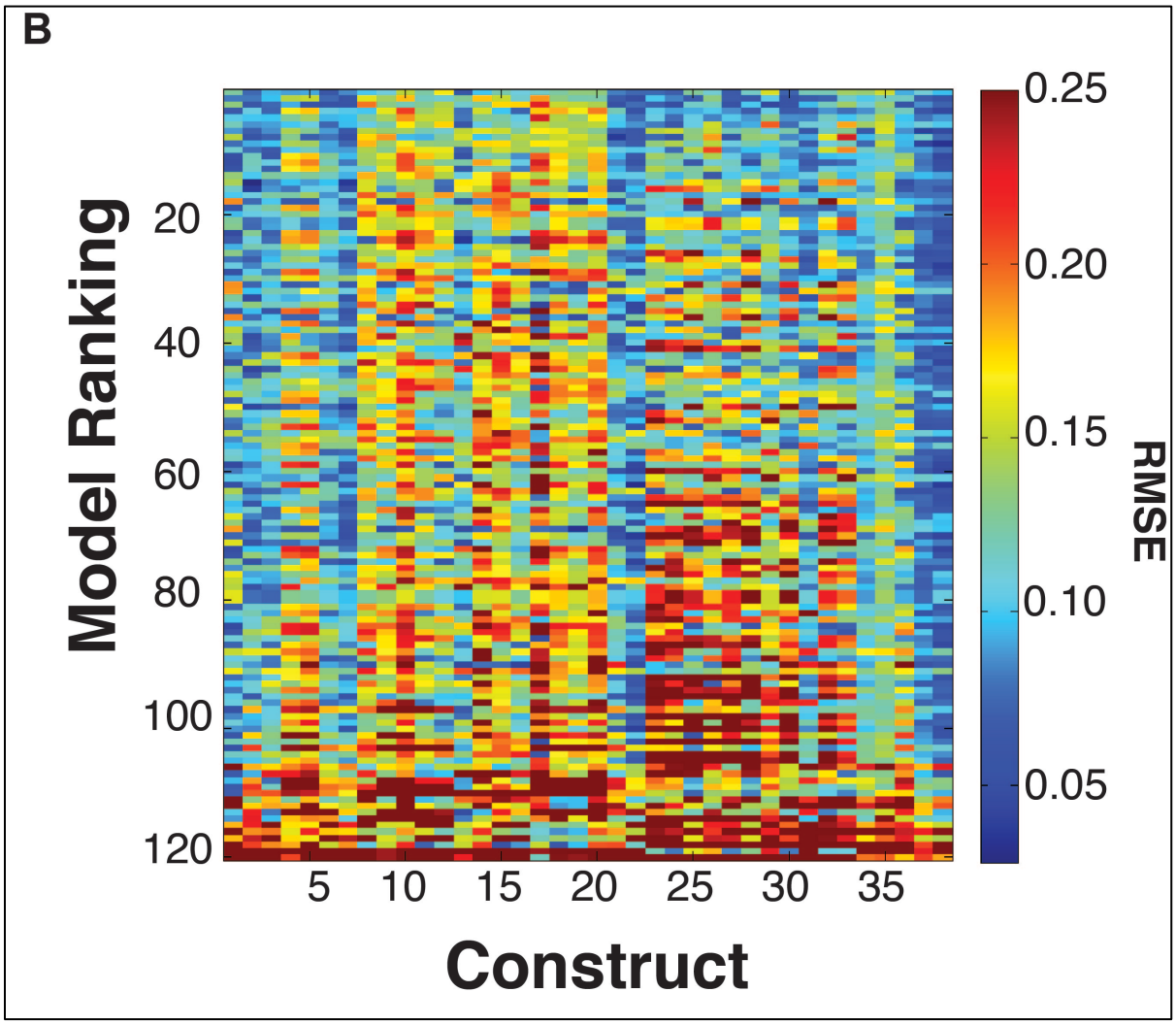


Figure S3.5 Performance of models on enhancer with relaxed PWM thresholds. (A) Heatmap of performance of 120 models on 38 *rho*NEE enhancer constructs with relaxed thresholds on the best PWM combinations. (B) Performance of 120 models on individual constructs (x-axis). Models are ranked along y-axis according to their performance on stringent PWM thresholds (see Fig. 3.3).

Figure S3.5 cont'd



Supplementary Table T1
Description and sequence information for enhancer constructs

Restriction sites used for cloning are capitalized, mutagenized binding site sequences are indicated in bold and capitalized letters.

Construct No.	Construct Name and Description	Sequence
1.	DMRW (wild-type rhoNEE from *D. mel*)	5' - ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcactcgcatatgttgagca catgttttgggggaaattcccgggcgacgggccaggaatcaacgtcc tgtcctgcgtgggaaaagcccacgtcctacccacgcccactcggttac ctGGCCGGCC – 3'
2.	DMRD1 (Dorsal1 site mutated)	5' – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggt**T TTAAAAACAC**acatcgcgaaacatttggcgcaacttgcggaa gacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctggga agcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggc gggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttga gcacatgttttgggggaaattcccgggcgacgggccaggaatcaac gtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccactcg gttacctGGCCGGCC – 3'
3.	DMRD2 (Dorsal2 site mutated)	5' – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg **CGTCAGTTAAT**gattcgcgatgccatgaggcactcgcatatgtt gagcacatgttttgggggaaattcccgggcgacgggccaggaatca acgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccact cggttacctGGCCGGCC – 3'
4.	DMRT1 (Twist1 site mutated)	5' – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcact**ACGCGTTGTT** gagcacatgttttgggggaaattcccgggcgacgggccaggaatca acgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccact

		cggttacctGGCCGGCC – 3’
5.	DMRT2 (Twist2 site mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcactcgcatatgttg**ACG CGTTGTT**ttgggggaaattcccgggcgacgggccaggaatcaa cgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccactc ggttacctGGCCGGCC – 3’
6.	DMRD3 (Dorsal3 site mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcactcgcatatgttgagca catgttttgg**TCTAGATTATC**gggcgacgggccaggaatcaa cgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccactc ggttacctGGCCGGCC – 3’
7.	DMRD4 (Dorsal4 site mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcactcgcatatgttgagca catgttttgggggaaattcccgggcgacgggccaggaatcaacgtcc tgtcctgcgt**AGGCCTGGTCA**acgtcctacccacgcccactcg gttacctGGCCGGCC – 3’
8.	DMRD1D2 (Dorsal1 and Dorsal2 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggt**T TTAAAAACAC**acatcgcgaaacatttggcgcaacttgcggaa gacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctggga agcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggc ggg**CGTCAGTTAAT**gattcgcgatgccatgaggcactcgcat atgttgagcacatgttttgggggaaattcccgggcgacgggccagga atcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcc cactcggttacctGGCCGGCC - 3’
9.	DMRD1T1 (Dorsal1 and Twist1 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtT **TTAAAAACAC**acatcgcgaaacatttggcgcaacttgcggaa gacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctggga agcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggc gggcggaatttcctgattcgcgatgccatgaggcact**ACGCGTT GTT**gagcacatgttttgggggaaattcccgggcgacgggccagga atcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcc cactcggttacctGGCCGGCC - 3’

10.	DMRD1T2 (Dorsal1 and Twist2 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggt**TTTAAAAACAC**acatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatgtt**ACGCGTTGTT**ttgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccactcggttacctGGCCGGCC - 3’
11.	DMRD1D3 (Dorsal1 and Dorsal3 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggt**TTTAAAAACAC**acatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttgagcacatgttttgg**TCTAGATTATC**gggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccactcggttacctGGCCGGCC – 3’
12.	DMRD1D4 (Dorsal1 and Dorsal4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcgg**tTTTAAAAACAC**acatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttgagcacatgttttgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtAGGCCTGGTCAacgtcctacccacgcccactcggttacctGGCCGGCC – 3’
13.	DMRD2T1 (Dorsal2 and Twist1 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcgggg**CGTCAGTTAAT**gattcgcgatgccatgaggcact**ACGCGTTGTT**gagcacatgttttgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgccccactcggttacctGGCCGGCC - 3’
14.	DMRD2T2 (Dorsal2 and Twist2 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcgggg**CGTCAGTTAAT**gattcgcgatgccatgaggcactcgcatatgtt g**ACGCGTTGTT**ttgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgccccactcggttacctGGCCGGCC - 3’
15.	DMRD2D3	5’ –

	(Dorsal2 and Dorsal3 sites mutated)	ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg **CGTCAGTTAAT**gattcgcgatgccatgaggcactcgcatatgtt gagcacatgttttgg**TCTAGATTATC**gggcgacgggccagga atcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcc cactcggttacctGGCCGGCC - 3’
16.	DMRD2D4 (Dorsal2 and Dorsal4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg **CGTCAGTTAAT**gattcgcgatgccatgaggcactcgcatatgtt gagcacatgttttgggggaaattcccgggcgacgggccaggaatca acgtcctgtcctgcgt**AGGCCTGGTCA**acgtcctacccacgcc cactcggttacctGGCCGGCC - 3’
17.	DMRT1T2 (Twist1 and Twist2 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcact**ACGCGTTGTT** g**ACGCGTTGTT**ttgggggaaattcccgggcgacgggccagg aatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgc ccactcggttacctGGCCGGCC - 3’
18.	DMRT1D3 (Twist1 and Dorsal3 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcact**ACGCGTTGTT** gagcacatgttttgg**TCTAGATTATC**gggcgacgggccagga atcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcc cactcggttacctGGCCGGCC – 3’
19.	DMRT1D4 (Twist1 and Dorsal4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcact**ACGCGTTGTT** gagcacatgttttgggggaaattcccgggcgacgggccaggaatca acgtcctgtcctgcgt**AGGCCTGGTCA**acgtcctacccacgcc cactcggttacctGGCCGGCC - 3’
20.	DMRT2D3 (Twist2 and	5- ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg

	Dorsal3 sites mutated)	ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcactcgcatatgttg**ACG CGTTGTT**ttgg**TCTAGATTATC**gggcgacgggccaggaa tcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgccc actcggttacctGGCCGGCC - 3’
21.	DMRT2D4 (Twist2 and Dorsal4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcactcgcatatgttg**ACG CGTTGTT**ttgggggaaattcccgggcgacgggccaggaatcaa cgtcctgtcctgcgt**AGGCCTGGTCA**acgtcctacccacgccc actcggttacctGGCCGGCC - 3’
22.	DMRD3D4 (Dorsal3 and Dorsal4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaaggacaccttgctgtgcggcgggaagcgcaagtggcggg cggaatttcctgattcgcgatgccatgaggcactcgcatatgttgagca catgttttgg**TCTAGATTATC**gggcgacgggccaggaatcaa cgtcctgtcctgcgt**AGGCCTGGTCA**acgtcctacccacgccc actcggttacctGGCCGGCC - 3’
23.	DMR4SM (4 Snail sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcg**CAGAGCTCGG**a agacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctggg aagcggaaaaagga**CAGGAGCTTG**tgcggcggga**ACGCC CGGCGG**cgggcggaatttcctgattcgcgatgccatgaggcactc gcatatgttga**GCATATGTTT**tgggggaaattcccgggcgacg ggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtccta cccacgcccactcggttacctGGCCGGCC - 3’
24.	DMRS1 (Snail2, Snail3 and Snail4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagac aagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc ggaaaaagga**CAGGAGCTTG**tgcggcggga**ACGCCG GCGG**cgggcggaatttcctgattcgcgatgccatgaggcactcgc atatgttga**GCATATGTTT**tgggggaaattcccgggcgacggg ccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacc cacgcccactcggttacctGGCCGGCC - 3’
25.	DMRS2 (Snail1, Snail3 and	5- ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcg**CAGAGCTCGG**a

	Snail4 sites mutated)	agacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcggga**ACGCCGGCGG**cgggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttga**GCATATGTTT**tgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctaccacgcccactcggttacctGGCCGGCC - 3’
26.	DMRS3 (Snail1, Snail2 and Snail4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcg**CAGAGCTCGG**aagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaagga**CAGGAGCTTG**tgcggcgggaacgcaagtggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttga**GCATATGTTT**tgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctaccacgcccactcggttacctGGCCGGCC - 3’
27.	DMRS4 (Snail1, Snail2 and Snail3 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcg**CAGAGCTCGG**aagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaagga**CAGGAGCTTG**tgcggcggga**ACGCCGGCGG**cgggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttgagcacatgttttgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccactcggttacctGGCCGGCC - 3’
28.	DMRS1S2 (Snail3 and Snail4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaaggacaccttgctgtgcggcggga**ACGCCGGCGG**cgggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttga**GCATATGTTT**tgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgccactcggttacctGGCCGGCC - 3’
29.	DMRS1S3 (Snail2 and Snail4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagcggaaaaagga**CAGGAGCTTG**tgcggcgggaacgcaagtggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttga**GCATATGTTT**tgggggaaattcccgggcgacgggccaggaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgccactcggttacctGGCCGGCC - 3’
30.	DMRS1S4 (Snail2 and Snail3 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtgggaaaaacacacatcgcgaaacatttggcgcaacttgcggaagacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctgggaagc

		ggaaaaagga**CAGGAGCTTG**tgcggcggga**ACGCCG GCGG**cgggcggaatttcctgattcgcgatgccatgaggcactcgc atatgttgagcacatgttttgggggaaattcccgggcgacgggccagg aatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgc ccactcggttacctGGCCGGCC - 3’
31.	DMRS2S3 (Snail1 and Snail4 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcg**CAGAGCTCGG**a agacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctggg aagcggaaaaaggacaccttgctgtgcggcgggaacgcaagtggc gggcggaatttcctgattcgcgatgccatgaggcactcgcatatgttga **GCATATGTTT**tgggggaaattcccgggcgacgggccaggaat caacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgccca ctcggttacctGGCCGGCC - 3’
32.	DMRS2S4 (Snail1 and Snail3 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcg**CAGAGCTCGG**a agacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctggg aagcggaaaaaggacaccttgctgtgcggcggga**ACGCCGG CGG**cgggcggaatttcctgattcgcgatgccatgaggcactcgcat atgttgagcacatgttttgggggaaattcccgggcgacgggccagga atcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcc cactcggttacctGGCCGGCC - 3’
33.	DMRS3S4 (Snail1 and Snail2 sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaacatttggcg**CAGAGCTCGG**a agacaagtgcggctgcaacaaaaagtcgcgaaacgaaactctggg aagcggaaaaaggga**CAGGAGCTTG**tgcggcgggaacgca agtggcgggcggaatttcctgattcgcgatgccatgaggcactcgcat atgttgagcacatgttttgggggaaattcccgggcgacgggccagga atcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcc cactcggttacctGGCCGGCC - 3’
34.	DMRB (Both bHLH sites mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaa**TGATTC**gcgcaacttgcggaa ga**TAGCGA**cggctgcaacaaaaagtcgcgaaacgaaactctg ggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagt ggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatg ttgagcacatgttttgggggaaattcccgggcgacgggccaggaatc aacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcccac tcggttacctGGCCGGCC - 3’
35.	DMRD1B (Both bHLH sites and Dorsal1 site mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggt**T TTAAAAACAC**acatcgcgaaa**TGATTC**gcgcaacttgcg gaaga**TAGCGA**cggctgcaacaaaaagtcgcgaaacgaaact ctgggaagcggaaaaaggacaccttgctgtgcggcgggaagcgca

		agtggcgggcggaatttcctgattcgcgatgccatgaggcactcgcat atgttgagcacatgttttgggggaaattcccgggcgacgggccagga atcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgcc cactcggttacctGGCCGGCC - 3’
36.	DMRBT2 (Both bHLH sites and Twist2 site mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaa**TGATTC**gcgcaacttgcggaa ga**TAGCGA**cggctgcaacaaaaagtcgcgaaacgaaactctg ggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagt ggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatg ttg**ACGCGTTGTT**ttgggggaaattcccgggcgacgggccag gaatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacg cccactcggttacctGGCCGGCC - 3’
37.	DMRBD3 (Both bHLH sites and Dorsal3 site mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaa**TGATTC**gcgcaacttgcggaa ga**TAGCGA**cggctgcaacaaaaagtcgcgaaacgaaactctg ggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagt ggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatg ttgagcacatgttttgg**TCTAGATTATC**gggcgacgggccagg aatcaacgtcctgtcctgcgtgggaaaagcccacgtcctacccacgc ccactcggttacctGGCCGGCC - 3’
38.	DMRBD4 (Both bHLH sites and Dorsal4 site mutated)	5’ – ACCGGTccttgggcaggatggaaaaatgggaaaacatgcggtg ggaaaaacacacatcgcgaaa**TGATTC**gcgcaacttgcggaa ga**TAGCGA**cggctgcaacaaaaagtcgcgaaacgaaactctg ggaagcggaaaaaggacaccttgctgtgcggcgggaagcgcaagt ggcgggcggaatttcctgattcgcgatgccatgaggcactcgcatatg ttgagcacatgttttgggggaaattcccgggcgacgggccaggaatc aacgtcctgtcctgcgt**AGGCCTGGTCA**acgtcctacccacgc ccactcggttacctGGCCGGCC - 3’
39.	DMBRKS (*D.mel brk* enhancer – 309 bp)	5’ – ACCGGTgggaaatccaaaacacaacccgagcccgatccttcgc tccttcgatttaagccaaagttagaggcacaggcacacatgtgtgtttg gtttgaacgggaaagccccattttaaagctggccaaccaacggcaa cacatgttcatgttaggaccgatacaggttgacattccctggaaggatg cacctctgggagattcccacaaccggcagcaggtcatgtccaaccg atcgttgcgggagccacttgtcccgaaaaaatccaaagaaactatca agtggcgtttagggaaactcaGGCCGGCC - 3’
40.	DMBRKL (*D.mel brk* enhancer – 649 bp)	5’ – ACCGGTaacaggtactacgatgatattggtcggaaaatacctgcg catcctggtggtttatggtgcggccgtaaatgcaagccaagttctttacg gcttctctggcacaaaccctaaatgtggattacgctaatattgcccccc ctaataaaaacggtcgttgtccagggccgaatattgcgtctgattggtttt tcccacgattacaattagccggacggacacaaactgacctgagctga cccgcaaaaagacacggttgtccggcagtcggaactgaaggaaac

		taaaggaaactgagggcaggtcagcgctatggattgtgcactaa gttgcttaatccgacgggaaatccaaaacacaacccgagcccgatc cttcgctccttcgatttaagccaaagttagaggcacaggcacacatgt gtgtttggtttgaacgggaaagccccattttaaagctggccaaccaac ggcaacacatgttcatgttaggaccgatacaggttgacattccctgga aggatgcacctctgggagattcccacaaccggcagcaggtcatgtcc aaccgatcgttgcgggagccacttgtcccgaaaaaatccaaagaaa ctatcaagtggcgtttagggaaactcaGGCCGGCC - 3’
41.	DMVNS (*D.mel vn* enhancer – 341 bp)	5’ – ACCGGTgggcatttcacttacctgcgtgggaaaatcgactaatctg cgaccgccccgaggagtcagtttttgtttttagagcggtaaaggacag gtaacgggccacatgtctggccggaaattccccgttgacccctgacc ccgtgtccttatgacgaattcgtcacttggcgtgagcacacctggatttc ccaccgcttagccagcggaaattccaaaacacctccggcccactgg ccctcaaaattgttatatgctctgctacgatgaagcagaagcagaagc agcagtgttttattggcggaagcatccgccaaattgcaccc aatctgcGGCCGGCC - 3’
42.	DMVNL (*D.mel vn* enhancer – 869 bp)	5’ – ACCGGTcaagttgagaaatttgcctttgatatcgaccacatgtgtgc acagcgaaaaaatattaggtggaaagttgaaaatattccgaattattc acaatatcatctgcaggattcttatttttaaaagcttccgatattacaaaa acactttcttggttgtaaatattttactgaatttgtgtaatttttctgtgtgcaa atatgcagccagttctggatcttccgaatcaccctgccttcgcgttttgca ccccgtcgctgtggagccatattttctttttagctgacgatttagtccatttc ccgctcataatcgcatgaagttgtttgcctccaccgaatggcttaatccg ccagatcgatgcgcctgtgttgactcaataattccctaacaactcttttta cgcattttattgaaagtgccgaagttagcgggcatttcacttacctgcgt gggaaaatcgactaatctgcgaccgccccgaggagtcagttttttgttttt agagcggtaaaggacaggtaacgggccacatgtctggccggaaatt ccccgttgacccctgaccccgtgtccttatgacgaattcgtcacttggcg tgagcacacctggatttcccaccgcttagccagcggaaattccaaaa cacctccggcccactggccctcaaaattgttatatgctctgctacgatg aagcagaagcagaagcagcagtgttttattggcggaagcatccgcc aaattgcacccaatctgcagtttgaagtgctcaaaacccccaccgctc ccctgtgaatttccgccggccggcaaggtgaccgtgtgctaaaacaa aatttttatatcgaaattgccgGGCCGGCC - 3’
43.	DMVNDS (*D.mel vnd* enhancer – 324 bp)	5’ – ACCGGTagaaattcccgtaggtgagagccagggaaaccccaat cgggaatgacatgtgtacgacagacatgggactcagatgccttcgag atactggcgtcacactgtctggcaatgggatttccgctcaggaggacg gggaatgcccgtgtagcctgtccatagcgtgggaaattcgcgagtcg gggtcttcgggaaaactcgaaatgggaaaaccggaagcaagcaa acttgcgccaacatgtggcacgacctgtttcgacccgtaaagagtccc tgctgacctgtgctgacctgcactgacccgaccaggtagGGCCG GCC - 3’
44.	DMVNDL	5’ –

	(*D.mel vnd* enhancer – 907 bp)	ACCGGTcaccctgcgagcctctgcctccatattagtgtttagatccc atatagctgatggacttgtccggctaatcctggtagctctttaaattaacc aagcgggcacagcgcgcaagtacaggacacagggtataattcccc gcctctttgatctggcaagtactcaaggtcctggcgaatggcggttggg aaattctggcttgttgttcgaccctggcttagaaattcccgtaggtgaga gccagggaaaccccaatcgggaatgacatgtgtacgacagacatg ggactcagatgccttcgagatactggcgtcacactgtctggcaatggg atttccgctcaggaggacggggaatgcccgtgtagcctgtccatagc gtgggaaattcgcgagtcggggtcttcgggaaaactcgaaatggga aaaccggaagcaagcaaacttgcgccaacatgtggcacgacctgtt tcgacccgtaaagagtccctgctgacctgtgctgacctgcactgaccc gaccaggtagctgcgatccttacgaggcggatttgcgtttaattgttgat ggtattaggcaaatcaaaactcggggtctgaccgggactaggtgtca ataatccagcgatttgggtgcacttattcaaagttaattccgggggaaa tgtgcgcgttttcggttccgaagcatgcctgcaggatgcacacccccc acctccttatcttcttaacaacggcaagtgcaaaaatctgtgaaagtca gagcgctacaggtagtgcaggtagtttcctttgcatatcccgaccaac agggacctccttttgttaaaccttccggccattcacacgattgacacag gatgtcgctgcaataagcatgaaacagggaaaaatcgGGCCGG CC - 3’
45.	DMVND1 (*D.mel vnd* enhancer – 362 bp)	5’ – ACCGGTgaccctggcttagaaattcccgtaggtgagagccaggg aaaccccaatcgggaatgacatgtgtacgacagacatgggactcag atgccttcgagatactggcgtcacactgtctggcaatgggatttccgctc aggaggacggggaatgcccgtgtagcctgtccatagcgtgggaaatt cgcgagtcggggtcttcgggaaaactcgaaatgggaaaaccggaa gcaagcaaacttgcgccaacatgtggcacgacctgtttcgacccgta aagagtccctgctgacctgtgctgacctgcactgacccgaccaggta gctgcgatccttacgaggcggatttgcgGGCCGGCC 3’
46.	DMSOGLS (*D.mel sog* enhancer – 406 bp)	5’ – ACCGGTgttgccaatgccattgcgcatacgccgtgtcgtctatatgg ctatatggctatatggctgtatggtgcggggaaatccccgtaatcgcag gtagaattccagccggtgccgaggcgggacctgctcgcacctctaat cccgccagggttttcgggacatgggatattcccgacggcacagcata gcactccgttttctttttttttttttattattattgtgtccagttttaatccggaaag cgggaattcccttccgctcgctgcctgcactgcgctgcgcagacgcat cggcgtccgtaagccgcttaccaaaaagatacgggtatacccaaat ggatgcctgcccatgtatatagaccattgggtggtatggaccatggac cataaagcGGCCGGCC - 3’
47.	DMSOGCE (*D.mel sog* enhancer – 564 bp)	5’ – ACCGGTtgtttatggcagccaattgatgccgactgacctgtgtgtgt gtgtgtgtgtgtgtggaagctcaggatggacagattcccgggtttcagc ggaacaggtaggctggtcgatcggaaattcccaccatacacatgtgg ctataatgccaacggcatcgaggtgcgaaaacagatgcagcctcat aaaaggggcgcagataaggtcgcggttgcgtgggaaaagcccatc cgaccaggaccaggacgaagcagtgcggttggcgcatcattgccgc

		catatctgctattcctacctgcgtggccatggcgatatccttgtgcaagg ataaggagcggggatcataaaacgctgtcgcttttgtttatgctgcttatt taaattggcttcttggcgggcgttgcaacctggtgctagtcccaatccc aatcccaattccaatccgtatacccgtatatccaatgcattctacctgtc ctgggaatttccgatttggccgcacccatatggccacggatgcgtgag agtgctctccgtgcgattctagatcatcGGCCGGCC - 3’
48.	DMTDP (*D.mel twi* enhancer – 1123 bp)	5’ – ACCGGTaagctcctaagtccaggtagttttgggacagggcaaaa ccctgttggtggtttttctaaggggaccatttcgagtcctgggttttgctatt acctaagccggcgatcggcgatctgcgatcggagatcttcgatcgtg gttttttccagcggaagttcgcgctctgcattaatcgggtatttttggtggc cccggcaggcaaacagataattatatccggaaatttgacttttcgctcg tatttttctggattttcggagctccgagccgcattcgcctgcgattttctcgg tacgtgtgtgtgggaattcactaattaggcataatgaaacctttcgtgg agttcccctcggttagggttgtggatttgcacgctttacgatggttggcaa ctaactgatgattatttaatagcggaatgatttcgatgggcgagcgtcta aacatttcggcttgtttcctgggaaattcctgcgatcccaaagtatatac aaatggaaaatcctcgcacagcaaagttgattgggtaaatatgcaat cagatagatataacttataatcgatttatatatacttacatctaaatgaatt tgatacccgatttatgtggatttttcgtgtttttgatcaggggagattcattgc gtcttatttttttttttacaaaaagaatataacttatttccatatcatttcaagtt tttaagttaaattcttgtctaaacctaaactaacgctgcaaacaacaac attcaatgaagtttactaaatgttcaattgggaattgcaaaacaaaatct tttctaccagaatgcaattttcaggaaatgcttttatgtaataaacataatt tatcattactctgaatgcacttttttcaaaacttagaaactctgtcctatgaa ttcccgtcgatccaaagatattctcaatccccttttgaatcaacaagtaa aatatttcaaaaattgccgacaattcccctcgtattccccgtcccgcatc ccaacacgcatacttcccaggcattttcccaaatcgagagaaaacccc aaagaataacccaagagaaacagaaaaatccagagcgtcgagtc aaggctctcttcaGGCCGGCC – 3’
49.	DMTDEPE (*D.mel twi* enhancer – 764 bp)	5’ – ACCGGTaagctcctaagtccaggtagttttgggacagggcaaaa ccctgttggtggtttttctaaggggaccatttcgagtcctgggttttgct attacctaagccggcgatcggcgatctgcgatcggagatcttcgatcg tggtttttttccagcggaagttcgcgctctgcattaatcgggtatttttggtg gccccggcaggcaaacagataattatatccggaaatttgacttttcgct cgtatttttctggattttcggagctccgagccgcattcgcctgcgattttct cggtacgtgtgtgtgggaattcactaattaggcataatgaaacctttcg tggagttcccctcggttagggttgtggatttgcacgctcctaggttaatta aattcaatgaagtttactaaatgttcaattgggaattgcaaaacaaaat ctttctaccagaatgcaattttcaggaaatgcttttatgtaataaacata atttatcattactctgaatgcacttttttcaaaacttagaaactctgtcctatg aattcccgtcgatccaaagatattctcaatccccttttgaatcaacaagt aaaatatttcaaaaattgccgacaattcccctcgtattccccgtcccgc atcccaacacgcatacttcccaggcattttcccaaatcgagagaaaa cccaaagaataacccaagagaaacagaaaaatccagagcgtcga

		gtcaaggctctcttcaGGCCGGCC – 3’
50.	DMTPE (*D.mel twi* enhancer – 318 bp)	5’ – ACCGGTgggaattgcaaaacaaaatcttttctaccagaatgcaatt ttcaggaaatgcttttatgtaataaacataatttatcattactctgaatgca ctttttcaaaacttagaaactctgtcctatgaattcccgtcgatccaaag atattctcaatccccttttgaatcaacaagtaaaatatttcaaaaattgc cgacaattcccctcgtattccccgtcccgcatcccaacacgcatacttc ccaggcattttcccaaatcgagagaaaacccaaagaataacccaa gagaaacagaaaaatccGGCCGGCC – 3’
51.	DMSPE (*D.mel sna* enhancer – 457 bp)	5’ – TTAATTAAggaggctgacatgcagactttgtacccggaaaaaca gacaagcccgcatagccaagtccgattttccgcgtcgtcaaaaaaaa aaaaccaggcgcggtggtaaaattggaaaaaatccatagcaaacc caaaagttttgataaacccaagaggtgacaaaaaattccagcgcgg aattccaattccccgcgatcctgcaaccaatggcgtgagggaagatg ggaaatgtcgagccaataacccagccgcatttccatttcgtatttccctg caatggattttccttcgacgaatgtccggatgtggatcagcgaaggattt aaagtaatttcccatcttatccaattacgcagtcagcgaaactcaagc aaaacaagccacgacgagtggaaaaactggctgggggttttctatat gtcctcacatcctacctggtgggtcGGCCGGCC – 3’
52.	DMSDP (*D.mel sna* enhancer – 867 bp)	5’ – ACCGGTaattgacaagaacaacaacaatgtctatggaaaatcga acttcatcccagcacctgcagaaatcccgagcgagtcggggaaaaa gtatttaacccccgaaagggttttccccaaaataatgaagtaatgaatg aagcggaaaacactggccgccaatctacctaatactaatgagcggg ccaacccgaccaggaatttttgcaagtcaggtacttcaacggatatat gggttcgacaagtgcggattttcccgcgacatcaatgaggacttggcc gggttatccgcggtgctcatcgggcaattccgcggccgaggacttcat cgtagtgatcattaggtagatatgtgcatggatgtgacatggcgatcatt gcgcggaataacacacgtaataaccgagatatccgggatgaccca ccaggtaggatgtgaggacatatagaaaacccccagccagtttttcc actcgtcgtggcttgttttgcttgagtttcgctgactgcgtaattggataa gatgggaaattactttaaatccttcgctgatccacatccggacattcgtc gaaggaaaatccattgcagggaaatacgaaatggaaatgcggctg ggttattggctcgacatttcccatcttccctcacgccattggttgcaggat cgcggggaattggaattccgcgctggaatttttgtcacctcttgggtttat caaaacttttgggtttgctatggatttttttccaattttaccaccgcgcctggt ttttttttttttgacgacgcggaaaatcggacttggctatgcgggcttgtctgt ttttccgggtacaaagtctgcatgtcagcctccGGCCGGCC – 3’
53.	DMSPEDE (*D.mel sna* enhancer – 805 bp)	5’ – ACCGGTaattgacaagaacaacaacaatgtctatggaaaatcga acttcatcccagcacctgcagaaatcccgagcgagtcggggaaaaa gtatttaacccccgaaagggttttccccaaaataatgaagtaatgaatg aagcggaaaacactggccgccaatctacctaatactaatgagcgg gccaacccgaccaggaatttttgcaagtcaggtacttcaacggatata tgggttcgacaagtgcggattttcccgcgacatcaatgaggacttggc

		cgggttatccgcggtgctcatcgggcaattccgcggccgaggacttca tcgtagtgatcattCCTAGGTTAATTAAgacccaccaggtag gatgtgaggacatatagaaaacccccagccagtttttccactcgtcgt ggcttgttttgcttgagtttcgctgactgcgtaattggataagatgggaaa ttactttaaatccttcgctgatccacatccggacattcgtcgaaggaaa atccattgcagggaaatacgaaatggaaatgcggctgggttattggct cgacatttcccatcttccctcacgccattggttgcaggatcgcggggaa ttggaattccgcgctggaatttttgtcacctcttgggtttatcaaaacttttg ggtttgctatggatttttttccaattttaccaccgcgcctggttttttttttttgac gacgcggaaaatcggacttggctatgcgggcttgtctgtttttccgggta caaagtctgcatgtcagcctccGGCCGGCC – 3’
54.	DERW (Putative *D.erecta* *rho*NEE – 261 bp)	5’ – ACCGGTgggaaaaacacacatcgcgaaacatttggcggaactt gcggaagacaagtgcggctgcaacaaaaaaaagtcgcgaaccaa aactctgggaagcggaaaaaggacaccttgctgtgcggcgggaag cgcaagtggcgggcggaatttcctgattcccggcccatgaggcactc gcatatgttgagcacatgttgggggaaattcccgggcgacgggccag gaatcaacgtcctgcgtgggaaaagccGGCCGGCC – 3’
55.	DARW (Putative *D.ananassa* *e rho*NEE – 328 bp)	5’ – ACCGGTgggaaaaacacacatcccgccgcacacacacacac atcgcgaaacatttggcggaacttggcggagagacaagtgccgcaa aacaaaaagtcgcgaaataaaactccgggaagtgaaaatgtgtca gagtgggaaaaggaggcaaggacacaggacatacaggacatac aggacaccttgctgtgcggcgtgtgggaaacaagttgcggaatttcct ggccaggaggcaagatagccgtcgcatatgttgagcacatgttggag gaaattcccgcacaagggccaggaatcaacgtccggcgtgggaaa agcccGGCCGGCC – 3’
56.	DMOJRW (Putative *D.mojavensi* *s rho*NEE – 455 bp)	5’ – ACCGGTgggaaaaccacactcacacacgtatacacacacagat acaaggactcacatagtttgtgggaactttcggacaagtgcaatacaa aagtcgaagtcgcgaaaacgcgttgagcaattcaaatgaaaatccg caatgcaacggaaggagcaaggacatcgcacatcgcacatcgca gaacctgcagcaatcttcctgtgcggaaattcctgaatcgcacatgtg gcacgcacatgttgctgctgcggcagtgggaaaaacgagacgaca aggaattccccgagagcagctcgccatgccacgcctacacgcccac acacccagcaaggcggcaattatgagtacctgtgactgcaacttgcg acttgcctcacctgaagtgtggaggccaaaaggtgaccgggacgtg cctcccagatttttgagagaacgtgggaaaaaagGGCCGGCC – 3’
57.	DGRW (Putative *D.erecta* *rho*NEE – 315 bp)	5’ – ACCGGTgggaaaaacacacatggactcacaggatgagcaatttt gttgacaagtgcaacaaaagtcgccgaaatcgcgaaatgcgcttca caatttcagatgaaaatccgcaatgcaacggaagggagcaaggac accgacgtagacgatgaacctgtgttcctgtcgcagcaacagcaaca aacaagtgcggaatttcctgactccggacacatgtggcacgcacatg tttggcagagcgaaaaagaaacacacatcaaaaaggaattcccca

		ggtcgagagagggaagcccacgcccacaggtgaGGCCGGC C – 3’
58.	DVRW (*D.virilis rho*NEE – 315 bp)	5’ – ACCGGTacgcacgcacacggcgatagaaattaacacgtagttta gcggaactttgtggcaagtgcaacaaaagtcgaagtcgcggacgatt caaatgaaaatctgcaatgctgcggaaggagcaaggacaacccac ctgtctatgagtgtgcgagtgtgcgagtgtgtgtgtgtgtgcgagtgtg tgtgcgtgcgtgtgtgtgcaacaagtgcggaaattcctgaatcgcacat gtggcacgcacatgtcgagcgggaaaaaaccgctcgatgctcggtc caaggaattccccgagccaagggaaGGCCGGCC – 3’
59.	DMTDE (*D.mel twi* enhancer – 333 bp)	5’ – ACCGGTggcaaaaccctgttggtggtttttctaaggggaccatttcg agtcctgggttttgctattacctaagccggcgatcggcgatctgcgatc ggagatcttcgatcgtggtttttttccagcggaagttcgcgctctgcattaa tcgggtatttttggtggccccggcaggcaaacagataattatatccgga aatttgacttttcgctcgtatttttctggattttcggagctccgagccgcatt cgcctgcgattttctcggtacgtgtgtgtgggaattcactaattagg cataatgaaacctttttcgtggagttccccGGCCGGCC – 3’

REFERENCES

REFERENCES

1. M. Levine, Transcriptional Enhancers in Animal Development and Evolution, *Current Biology* **20**, R754–R763 (2010).

2. P. J. Wittkopp, G. Kalay, Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence, *Nat Rev Genet* **13**, 59–69 (2012).

3. M. Bulger, M. Groudine, Functional and Mechanistic Diversity of Distal Transcription Enhancers, *Cell* **144**, 327–339 (2011).

4. L. Dunipace, A. Ozdemir, A. Stathopoulos, Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression, *Development* **138**, 4075–4084 (2011).

5. S. Barolo, Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy, *Bioessays* **34**, 135–141 (2011).

6. M. W. Perry, A. N. Boettiger, M. Levine, Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo, *Proc Natl Acad Sci U S A* **108**, 13570–13575 (2011).

7. N. Frankel *et al.*, Phenotypic robustness conferred by apparently redundant transcriptional enhancers, *Nature* **466**, 490–493 (2010).

8. H. N. Cai, D. N. Arnosti, M. Levine, Long-range repression in the Drosophila embryo, *Proc Natl Acad Sci U S A* **93**, 9309–9314 (1996).

9. T. K. Kim, T. Maniatis, The mechanism of transcriptional synergy of an in vitro assembled interferon-beta enhanceosome, *Molecular Cell* **1**, 119–129 (1997).

10. S. D. Hanes, G. Riddihough, D. Ish-Horowicz, R. Brent, Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen, *Molecular and Cellular Biology* **14**, 3364–3375 (1994).

11. V. J. Makeev, A. P. Lifanov, A. G. Nazina, D. A. Papatsenko, Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information, *Nucleic Acids Research* **31**, 6016–6026 (2003).

12. M. M. Kulkarni, D. N. Arnosti, cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster, *Molecular and Cellular Biology* **25**, 3411–3420 (2005).

13. E. E. Hare, B. K. Peterson, V. N. Iyer, R. Meier, M. B. Eisen, Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation, *PLoS Genetics* **4**, e1000106 (2008).

14. C. I. Swanson, N. C. Evans, S. Barolo, Structural Rules and Complex Regulatory

Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer, *Developmental Cell* **18**, 359–370 (2010).

15. D. May *et al.*, Large-scale discovery of enhancers from human heart tissue, *Nat Genet* **44**, 89–93 (2012).

16. A. Rada-Iglesias *et al.*, A unique chromatin signature uncovers early developmental enhancers in humans, *Nature* **470**, 279–283 (2010).

17. B. P. Berman *et al.*, Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome, *Proceedings of the National Academy of Sciences of the United States of America* **99**, 757–762 (2002).

18. A. Ay, D. N. Arnosti, Mathematical modeling of gene expression: a guide for the perplexed biologist, *Critical Reviews in Biochemistry and Molecular Biology* **46**, 137–151 (2011).

19. L. Bintu *et al.*, Transcriptional regulation by the numbers: models, *Current Opinion in Genetics & Development* **15**, 116–124 (2005).

20. W. D. Fakhouri *et al.*, Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo, *Mol. Syst. Biol.* **6**, 341 (2010).

21. H. Janssens *et al.*, Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene, *Nat Genet* **38**, 1159–1165 (2006).

22. E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, U. Gaul, Predicting expression patterns from regulatory sequence in Drosophila segmentation, *Nature* **451**, 535–540 (2008).

23. X. He, M. A. H. Samee, C. Blatti, S. Sinha, Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression, *PLoS Comp Biol* **6**, e1000935 (2010).

24. O. Rubel *et al.*, *Pointcloudxplore: Visual analysis of 3d gene expression data using physical views and parallel coordinates. Eurographics* (IEEE-VGTC Symposium on Visualization, 2006), pp. 11–12 May– 2006 Braga– Portugal.

25. S. V. E. Keränen *et al.*, Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution II: dynamics, *Genome Biol* **7**, R124 (2006).

26. C. L. L. Hendriks *et al.*, Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline, *Genome Biol* **7**, R123–R123 (2006).

27. X.-Y. Li *et al.*, Transcription factors bind thousands of active and inactive regions in

the Drosophila blastoderm, *PLoS Biol* **6**, e27–e27 (2008).

28. S. MacArthur *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–R80 (2009).

29. T. Sandmann *et al.*, A core transcriptional network for early mesoderm development in Drosophila melanogaster, *Genes & Development* **21**, 436–449 (2007).

30. J. Zeitlinger *et al.*, Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo, *Genes & Development* **21**, 385–390 (2007).

31. Q. He *et al.*, High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species, *Nat Genet* **43**, 414–420 (2011).

32. M. D. Biggin, Animal Transcription Networks as Highly Connected, Quantitative Continua, *Developmental Cell* **21**, 611–626 (2011).

33. R. P. Zinzen, K. Senger, M. Levine, D. Papatsenko, Computational models for neurogenic gene expression in the Drosophila embryo, *Curr. Biol.* **16**, 1358–1365 (2006).

34. J. Gertz, E. D. Siggia, B. A. Cohen, Analysis of combinatorial cis-regulation in synthetic and genomic promoters, *Nature* **457**, 215–218 (2008).

35. Y. T. Ip, R. E. Park, D. Kosman, E. Bier, M. Levine, The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo, *Genes & Development* **6**, 1728–1739 (1992).

36. R. Sayal, S.-M. Ryu, D. N. Arnosti, Optimization of reporter gene architecture for quantitative measurements of gene expression in the Drosophila embryo, *Fly (Austin)* **5**, 47–52 (2011).

37. J. Jiang, D. Kosman, Y. T. Ip, M. Levine, The dorsal morphogen gradient regulates the mesoderm determinant twist in early Drosophila embryos, *Genes & Development* **5**, 1881–1891 (1991).

38. D. Papatsenko, ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors, *Bioinformatics* **23**, 1032–1034 (2007).

39. S. Small, A. Blair, M. Levine, Regulation of even-skipped stripe 2 in the Drosophila embryo, *EMBO J.* **11**, 4047–4057 (1992).

40. H. Janssens *et al.*, A high-throughput method for quantifying gene expression data from early Drosophila embryos, *Dev. Genes Evol.* **215**, 374–381 (2005).

41. D. Kosman *et al.*, Multiplex detection of RNA expression in Drosophila embryos,

Science **305**, 846 (2004).

42. A. Ay, W. D. Fakhouri, C. Chiu, D. N. Arnosti, Image processing and analysis for quantifying gene expression from early Drosophila embryos, *Tissue Eng Part A* **14**, 1517–1526 (2008).

43. E. Myasnikova, M. Samsonova, D. Kosman, J. Reinitz, Removal of background signal from in situ data on the expression of segmentation genes in Drosophila, *Dev. Genes Evol.* **215**, 320–326 (2005).

44. C. M. Bergman, J. W. Carlson, S. E. Celniker, Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster, *Bioinformatics* **21**, 1747–1749 (2005).

45. M. B. Noyes *et al.*, A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system, *Nucleic Acids Research* **36**, 2547–2560 (2008).

46. A. Ozdemir *et al.*, High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation, *Genome Res.* **21**, 566–577 (2011).

47. T. L. Bailey *et al.*, MEME SUITE: tools for motif discovery and searching, *Nucleic Acids Research* **37**, W202–W208 (2009).

48. N. Hansen, S. D. Müller, P. Koumoutsakos, Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), *Evol Comput* **11**, 1–18 (2003).

49. Y. Suleimenov *et al.*, Global Parameter Estimation for Thermodynamic Models of Transcriptional Regulation,(submitted, 2012).

50. M. Matsumoto, T. Nishimura, Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Trans. Model. Comput. Simul.* **8**, 3–30 (1998).

51. J. M. Dresch, X. Liu, D. N. Arnosti, A. Ay, Thermodynamic modeling of transcription: sensitivity analysis differentiates biological mechanism from mathematical model-induced effects, *BMC Syst Biol* **4**, 142–142 (2010).

52. P. Tomancak *et al.*, Global analysis of patterns of gene expression during Drosophila embryogenesis, *Genome Biol* **8**, R145 (2007).

53. S. Kumar *et al.*, FlyExpress: visual mining of spatiotemporal patterns for genes and publications in Drosophila embryogenesis, *Bioinformatics* **27**, 3319–3320 (2011).

54. E. Bier, L. Y. Jan, Y. N. Jan, rhomboid, a gene required for dorsoventral axis establishment and peripheral nervous system development in Drosophila melanogaster,

Genes & Development **4**, 190–203 (1990).

55. J. Crocker, Y. Tamori, A. Erives, Evolution acts on enhancer organization to fine-tune gradient threshold readouts, *PLoS Biol* **6**, e263 (2009).

56. G. T. Reeves, A. Stathopoulos, Graded Dorsal and Differential Gene Regulation in the Drosophila Embryo, *Cold Spring Harbor Perspectives in Biology* **1**, a000836–a000836 (2009).

57. P. Szymanski, M. Levine, Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo, *EMBO J.* **14**, 2229–2238 (1995).

58. S. Gonzalez-Crespo, M. Levine, Interactions between dorsal and helix-loop-helix proteins initiate the differentiation of the embryonic mesoderm and neuroectoderm in Drosophila, *Genes & Development* **7**, 1703–1713 (1993).

59. S. Gray, P. Szymanski, M. Levine, Short-range repression permits multiple enhancers to function autonomously within a complex promoter, *Genes & Development* **8**, 1829–1838 (1994).

60. M. Tompa *et al.*, Assessing computational tools for the discovery of transcription factor binding sites, *Nature Biotechnology* **23**, 137–144 (2005).

61. J. Jiang, M. Levine, Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen, *Cell* **72**, 741–752 (1993).

62. H.-L. Liang *et al.*, The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila, *Nature* **456**, 400–403 (2008).

63. C.-Y. Nien *et al.*, Temporal coordination of gene networks by Zelda in the early Drosophila embryo, *PLoS Genetics* **7**, e1002339 (2011).

64. L. M. Liberman, A. Stathopoulos, Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence, *Developmental Biology* **327**, 578–589 (2009).

65. G. Junion *et al.*, A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History, *Cell* **148**, 473–486 (2012).

66. M. S. Halfon, Y. Grad, G. M. Church, A. M. Michelson, Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model, *Genome Res.* **12**, 1019–1028 (2002).

67. S. Aerts *et al.*, Robust Target Gene Discovery through Transcriptome Perturbations and Genome-Wide Enhancer Predictions in Drosophila Uncovers a Regulatory Basis for Sensory Specification, *PLoS Biol* **8**, e1000435 (2010).

68. R. N. Gutenkunst *et al.*, Universally sloppy parameter sensitivities in systems biology models, *PLoS Comp Biol* **3**, 1871–1878 (2007).

69. L. A. Mirny, Nucleosome-mediated cooperativity between transcription factors, *Proc Natl Acad Sci U S A* **107**, 22534–22539 (2010).

70. L. M. Li, D. N. Arnosti, Long- and short-range transcriptional repressors induce distinct chromatin States on repressed genes, *Current Biology* **21**, 406–412 (2011).

71. J. Cowden, M. Levine, The Snail repressor positions Notch signaling in the Drosophila embryo, *Development* **129**, 1785–1793 (2002).

72. V. Morel, F. Schweisguth, Repression by suppressor of hairless and activation by Notch are required to define a single row of single-minded expressing cells in the Drosophila embryo, *Genes & Development* **14**, 377–388 (2000).

73. J. Svaren, Transcription factors vs nucleosomes: regulation of the PH05 promoter in yeast, *Trends in Biochemical Sciences* **22**, 93–97 (1997).

74. L. Bai, A. V. Morozov, Gene regulation by nucleosome positioning, *Trends Genet* **26**, 476–483 (2010).

75. G. O. Bryant *et al.*, Activator control of nucleosome occupancy in activation and repression of transcription, *PLoS Biol* **6**, 2928–2939 (2008).

CHAPTER IV

Conclusions and Future Perspectives

Conclusions

This thesis is focused on developing a quantitative, predictive model of enhancer function in metazoans. Studies presented here provide new insights on enhancer structure and evolution as well as present a framework for quantitative studies of enhancers. Now I will discuss the major conclusions of this thesis, as well as provide some perspective on outstanding questions in the field.

Individual activator binding sites contribute quantitatively to enhancer function

rhomboid enhancer contains four Dorsal, two Twist and four Snail binding sites. Analysis of mutations of one and two activator sites in all possible combinations revealed that loss of any single site led to similar quantitative effects, while mutation of any two sites had a highly variable effect on gene expression. These observations imply that there is a complex landscape for positioning of individual activator binding sites. Previous studies have proposed that Dorsal and Twist activators show two modes of transcriptional synergy – one requiring linked sites for cooperative binding, and another via separate interactions with different components of basal machinery for effective recruitment, which don't require linked sites (*1*). It is likely that the transcriptional effects seen here are the results of both of these processes. This feature has implications for the structure of enhancers regulated by these same sets of proteins, as well as other similar systems where multiple activator bind cooperatively to DNA and interact with basal transcriptional machinery to regulate gene expression.

Repressor sites are positioned strategically on enhancer for effective repression

Snail is a short-range transcriptional repressor, viz., it can repress activator when they bind within 100 bp of a Snail site. My analysis of mutation of single and multiple

Snail sites showed that proximity to multiple activator sites is a crucial factor to determine effective Snail repression. Snail 2 and Snail 3 sites, which are in close proximity to at least four activator sites, were the most effective. These results are in agreement with previous studies on analysis of Snail sites in minimal *rho* enhancer (*2*), and establish how these features of cooperative activation and short-range repression can influence enhancer structure.

Repressors function by recruiting different corepressor proteins to enhancers and establishing repressive chromatin. Studies on chromatin effects of short-range repressors have shown that these repressors cause enhancer-wide decrease in histone acetylation, which is a histone modification associated with active enhancers. Additionally, repressors also reduce activator occupancy on enhancers (*3*). It remains to be seen how chromatin effects mediated by repressors compare with number and placement of repressor sites in an enhancer.

Parameter estimation can provide biochemical insights on enhancer function

My studies on *rhomboid* enhancer constructs, coupled with parameter estimation and thermodynamic modeling, set out to determine the collection of parameters that can fit the dataset. Three categories of parameters were estimated – protein cooperativity, repressor quenching and protein scaling factor. Scaling factor for a protein is an all-encompassing term for the different biochemical activities mediated by a protein that influence transcription positively or negatively. Several models that fit the dataset quite well came up with highly similar parameter values. Dorsal-Dorsal, Twist-Twist and Dorsal-Twist cooperativity came up frequently with positive values, with little change over distance. These values imply that these activator sites function synergistically.

Dorsal-Twist cooperativity has also been shown to be important for neurectoderm expression in previous studies. Positive cooperativity among all activator sites is in agreement with my observations that mutation of any two activator sites produced large, deleterious effects on activation. In contrast, Snail-Snail cooperativity values came up frequently negative, implying that Snail works in an additive fashion on the enhancer for effective repression.

Scaling factors for Dorsal and Twist were generally very low and equal for both the factors, indicating that these activators have equally positive effects on gene expression. In line these observations, previous studies have noted that synthetic enhancers with arrays of multiple Dorsal sites and Twist sites are less effective than enhancers containing both Dorsal and Twist sites, indicating low activation efficiency of a single class of activator and high transcriptional synergy between Dorsal and Twist (*1*).

Quenching parameters were estimated separately for Snail on Dorsal and Twist. These parameters were meant to capture distance-dependent efficiency of repression by Snail on two activators. Surprisingly, the trends were remarkably different for Dorsal and Twist, implying that Snail represses these activators by two different mechanisms. Quenching of Dorsal by Snail was estimated as a monotonically decreasing function. For Twist, it was either very low for all distances or low for proximal distances and high at intermediate distances. Snail recruits a variety of corepressors for its repression activity. It is likely that different corepressors are required to inhibit different activators, which may be reflected in these parameter trends.

Thermodynamic models can predict the output of diverse regulatory elements

The purpose of thermodynamic modeling is to understand enhancer function by enumerating all the binding events occurring on an enhancer and correlating them to gene expression. *rho* enhancer is one of the several enhancers regulated by Dorsal, Twist and Snail in early fly embryo. The reason for generating a quantitative dataset was to train the model on this well-characterized enhancer to predict the output of other regulatory elements in the regulatory network.

Several top-ranked models incorporating diverse cooperativity and quenching functions were able to successfully predict the output of other patterning enhancers. It is interesting to note that both the gene expression patterns as well as landscape of binding sites on these enhancers is widely divergent from *rho*. Still, parameters obtained from modeling on *rho* were able to predict quantitatively the output of these enhancers.

Although this is encouraging, there were some enhancers where most of the models performed poorly. Most notably, gene activation driven by mesodermal enhancers was not predicted well. Previous studies have indicated that linked activator binding sites are important for neurectoderm expression, where Dorsal and Twist gradients fall sharply. My study was focused on “model learning” from a neurectodermal enhancer and applying those insights to mesodermal enhancers. It is likely that parameters of activation are different for mesodermal and neurectodermal enhancers, which can be investigated by incorporating sufficient data from mesodermal enhancers into modeling.

Future Perspectives

Henceforth, I would like to discuss how further studies may enhance our knowledge of enhancer structure and function.

Extension of quantitative modeling to other systems

The study presented here built upon a large amount of system information and a vast body of knowledge for a well-characterized enhancer, making it an ideal "test-bed" for quantitative studies. The successes and failures of this model illustrate the necessary requirements for application of thermodynamic models to other systems, as well as guidelines to ascertain the size of datasets for the purposes of modeling.

First, the structure of genetic network involved in regulation of target gene expression should be well-characterized. The proteins and signals active in regulating gene expression constitute important inputs to the system. If the proteins are DNA-binding transcription factors, their binding to regulatory sequences and contribution to transcriptional regulation can be estimated using thermodynamic models. Next, it is important to have quantitative information about relative levels of transcription factors, if enhancers are responsive to these differences. One of the most important components of mathematical modeling is the generation of quantitative, systematic perturbation datasets for the system of interest. The perturbation can be of various kinds – gene dosage effects, protein-binding site mutants, effects of inducers, etc. A sufficiently large dataset provides a spectrum of system-wide responses for reliable parameter estimation. It is also important to characterize the regulatory sequences involved in transcriptional regulation of a particular gene of interest, in terms of both annotation of minimal enhancer sequences sufficient to drive gene expression when tested in reporter gene assays, as well as in terms of identity of transcription factors binding to enhancers and the sequences bound by them. In this respect, a researcher can take advantage of availability of large, genome-wide datasets (e.g., ENCODE) of binding profiles of

various transcription factors in different developmental contexts and environmental conditions.

Effect of chromatin context on enhancer function

The packaging of DNA into nucleosomes in eukaryotes presents a problem for regulation of gene expression. How do transcription factors get access to regulatory elements? A workaround to this problem has been the use of several cofactors by transcription factors that can modify histone tails to make chromatin compact or loose, as well as displace nucleosomes using ATP-dependent chromatin remodeling enzyme complexes to expose binding sites for other proteins. Recent genome-wide nucleosome positioning maps provide us with a static snapshot of their locations. It is necessary to integrate this information with knowledge gleaned from experiments documenting removal/remodeling of nucleosomes by cofactors recruited by various transcription factors into future modeling efforts.

Recent studies have provided detailed gene-specific as well as genome-wide maps of chromatin modifications catalyzed by various enzymes recruited by activator and repressor proteins. Biochemical experiments are providing us with detailed insights about how these modifications are dynamically “written” and “erased” from chromatin. Additionally, specific chromatin signatures have been found to correlate with temporal shifts in enhancer activity (*4*). In order to assess enhancer activity in terms of temporal dynamics, it is important to incorporate tissue- and stage-specific chromatin-level information into thermodynamic models.

REFERENCES

REFERENCES

1. P. Szymanski, M. Levine, Multiple modes of dorsal-bHLH transcriptional synergy in the Drosophila embryo, *EMBO J.* **14**, 2229–2238 (1995).

2. S. Gray, P. Szymanski, M. Levine, Short-range repression permits multiple enhancers to function autonomously within a complex promoter, *Genes & Development* **8**, 1829–1838 (1994).

3. L. M. Li, D. N. Arnosti, Long- and short-range transcriptional repressors induce distinct chromatin States on repressed genes, *Current Biology* **21**, 406–412 (2011).

4. S. Bonn *et al.*, Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development, *Nat Genet* **44**, 148–156 (2012).