UNDERSTANDING CO-EXPRESSED GENE SETS BY IDENTIFYING REGULATORS AND MODELING GENOMIC ELEMENTS

BY

CHARLES A BLATTI III

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Associate Professor Saurabh Sinha, Chair
Assistant Professor Jian Ma
Professor Gene Robinson
Associate Professor Scot Wolfe, University of Massachusetts Medical School
Professor ChengXiang Zhai

ABSTRACT

Genomic researchers commonly study complex phenotypes by identifying experimentally derived sets of functionally related genes with similar transcriptional profiles. These gene sets are then frequently subjected to statistical tests of association relating them to previously characterized gene sets from literature and public databases. However, few tools exist examining the non-coding, regulatory sequence of gene sets for evidence of a shared regulatory signature that may signal the involvement of important DNA-binding proteins called transcription factors (TFs). Here, we proposed and developed new computational methods for identifying major regulatory features of co-expressed gene sets that incorporate TF-DNA binding specificities ("motifs") with other important features such as sequence conservation and chromatin structure. We additionally demonstrated a novel approach for discovering regulatory signatures that are shared across gene sets from multiple experimental conditions or tissues. Given the co-expressed genes of a particular cell type, we also attempted to annotate their specific regulatory sequences ("enhancers") by constructing models of enhancer activity that incorporate the expression and binding specificities of the relevant transcription factors. We first developed and tested these models in well-characterized cell types, and then evaluated the extent to which these models were applicable using only minimal experimental evidence to poorly characterized systems without known transcriptional regulators and functional enhancers. Finally, we developed a network-based algorithm for examining novel gene sets that integrates many diverse types of biological evidences and relationships to better discover functionally related genes. This novel approach processed a comprehensive, heterogeneous network of biological knowledge and ranked genes and molecular properties represented in the network for their relevance to the given set of co-expressed genes.

ACKNOWLEDGMENTS

First and foremost, I am grateful to my advisor, Professor Saurabh Sinha. His enthusiasm, guidance, and support have fostered my ability to conduct research and have provided energy and ideas into the methods developed and insights obtained in this dissertation. His personable manner and straightforward ability to explain have been invaluable to me when I have had to grapple with complex issues in biology, computer science, or career development. I thank him and the Department of Computer Science for funding me and my research throughout the years.

I would also like to profusely thank Majid Kazemian, my fellow graduate student and collaborator on many of the projects presented in this dissertation. Working with Majid is always an enjoyable and enlightening experience, and his insights and contributions have been invaluable to the quality of this work and the development of my critical thinking and programming abilities. Most importantly, I value the time and lessons he provided to me on being a true, great friend.

A debt of thanks is also owed to the other collaborators of the research efforts presented here. Several of my research projects were made more impactful because of the involvement and support of Michael Brodsky and Scot Wolfe. Additionally, working with Gene Robinson and Seth Ament provided an excellent experience that improved my ability to rigorously justify methods and results.

My research was funded several times in the past few years by the Debra and Ira Cohen Fellowship. Mr. and Mrs. Cohen are inspirationally full of generosity, alumni pride, and a sincere interest in the advancement of science. I am not able to thank them enough for their support.

I thank the remaining members of my committee, Jian Ma and ChengXiang Zhai, for their guidance and support. I send my gratitude to the present and former students I have worked with in the Sinha lab, in CS, and at the IGB for discussing research issues with me while I was working and for sharing their company and laughs when I was not.

I am very grateful for my long-time friends, who continue to tolerate me and enrich my life with fun and companionship. I thank my entire family, especially my sisters for their love. Most of all, I thank my parents. The lessons they taught me, intentional and not, about leading a good life worthy of their pride have propelled me to this point and will guide me always.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

A major paradigm of genomic research is for investigators to experimentally identify sets of co-expressed genes in their system of interest. Researchers strive to uncover insights of the system based on characterization of these novel gene sets. There are many methods to characterize the expression of genes in tissues and other cell types, including measurement by *in situ* hybridization techniques [1], microarrays [2], or high throughput sequencing technologies like RNA-seq [3]. These methods have been employed to find sets of genes that are naturally expressed in different tissues, from developmental cell types [1] to regions of the adult brain [4]. They are able to determine differentially expressed genes in tissues in response to chemical stimuli [5], in regulatory networks that are affected by the disturbance (knockdown) of an important regulator [6], or in the brains of social animals that are exposed to behavioral provocations [7]. Important to the study of human health, these experimental assays find sets of genes whose transcription has been disrupted in the transition from healthy to cancerous tissues [8].

Once a novel, experimentally characterized gene set is identified, it is primarily analyzed by comparing it to other curated or experimental gene sets [9]. For example, researchers would like to understand if their novel gene set is enriched with genes that direct a particular biological process, perform a specific molecular function, are part of the same cellular component [10], or catalyze specific metabolic pathways [11]. Investigators are also interested in identifying similar gene sets from other experimental conditions and tissues. For example, they may want to know if their set of differentially expressed genes from a metastatic tissue originating from a breast cancer primary tumor is more similar to genes identified in other breast cancer tumors or in other metastatic tumors from different cancers. There are many published tools [9] that perform enrichment analysis on experimentally produced, co-expressed gene sets with two of the most popular being DAVID [12] and GSEA [13].

1.1 Predicting transcriptional regulators of a co-expressed gene set

One important reason for many genes to have a common transcriptomic profile is that they may share regulatory signals. Proteins that bind to the DNA near a gene and affect its level of expression are called transcription factors (TFs). The nearby regulatory sequence of genes of a novel co-expressed set may be enriched in binding sites of the same TF [14]. These major regulatory proteins are of great interest to investigators, but fewer tools exist to identify them.

There are some tools [15, 16] that rely on experimentally characterized TF binding from chromatin immunoprecipitation (ChIP) based methods [17] to identify these regulators, but producing such 'ChIP-seq' data for hundreds of TFs is currently infeasible. Other methods search the regulatory DNA of the co-expressed genes for overrepresented sequence patterns [18, 19], but these suffer from poor statistical power [20].

In Chapter 3, we will present a method to find regulatory signal enrichments by approximating ChIP TF binding information using the DNA binding specificity ("motif") of a transcription factor. Unlike the above-mentioned approaches, this approach allows us to specifically test for enrichment with hundreds of potential regulators. To distinguish our method from other motif-based enrichment tools [21, 22], we developed procedures from incorporating TF binding conservation and information on chromatin structure [23]. We developed and evaluated our regulatory enrichment tool in ~200 co-expressed gene sets from embryonic development in the fruit fly [1] and compared our method to alternatives. From this analysis, we built a compendium of >1000 relationships between these gene sets and their predicted major regulators. This compendium enabled us to discover additional biological insights into the TFs and cell types involved in this developmental system.

1.2 Discovery of shared regulatory signatures across multiple gene sets

Often, researchers want to analyze multiple co-expressed gene sets from several related experiments. Examples of this include gene sets from separate studies of a type of cancer [8], from different brain tissues of organisms exhibiting the same behavior [24, 25], or from orthologous tissues across several species [7]. The most common approach to examining multiple gene sets at once is to find core gene modules with biclustering tools [26, 27]. These core modules are a subset of genes that share a particular expression pattern across several, but not necessarily all of the examined sets. After finding the core modules, sequence patterns are found in the regulatory sequence of the module genes [18, 28]. This is sometimes done in an iterative manner that converges on gene modules with the strongest sequence signals [29].

The approach we present in Chapter 4 is distinct from the methods discussed above. We independently searched for the regulatory signals in each of the user-provided gene sets and then combined the significance of those signals to find common regulators. Rather than applying the standard techniques for combining significance values [30, 31], we developed a novel test statistic that enables us to identify instances when the regulator is important in only a subset of

the original gene sets. We also introduced a method for identifying shared *combinations* of regulatory signals because TFs are often observed interacting during transcriptional regulation in eukaryotes. We evaluated and compared our method, called '*cis*-Metalysis', on synthetic data and in the context of differentially regulated gene sets from eleven determinants of honeybee maturation. Finally, we applied our novel tool to gene sets derived from human cancer tissues and the brains of aggressively behaving social animals [7].

1.3 Modeling enhancers of gene expression in well and poorly studied cell types

Sets of genes with the same expression pattern in a cell type are likely to be regulated by the shared regulators that are expressed in the cell type. These genes are also expected to contain regulatory control regions (enhancers) that encode binding sites for the relevant TFs. These enhancers are typically 500-1000 base pair sequences that directly affect the transcription of the genes [32]. The annotation of regulatory enhancers enables researchers to better understand the signals and mechanisms that affect specific transcriptional responses. A collection of annotated enhancers also provides an important subset of functional segments in the large non-coding genome that may aid in the discovery of genetic mutations that correlate with disease [33]. There are many methods that annotate putative enhancers that rely only on DNA sequence and/or TF motifs [34-36]. There are also methods that rely on experimental assays for characterizing structural or state information of chromatin to identify potential regulatory sequences [23, 37]. For a cell type with a set of experimentally validated enhancers and a collection of known regulatory TFs, models of enhancer activity can be developed and applied to enhancer annotation. These models incorporate the binding and expression information of the relevant TFs to predict the gene expression driven by the enhancers of the cell type. Such models have been developed with ChIP [38-42] and motif [43-48] based TF binding features and using Bayesian Network [49], support vector machine [38], and thermodynamic [44, 45] modeling frameworks.

Chapter 5 is dedicated to the construction of simple models of enhancer activity and their application in annotating regulatory sequences for genes expressed in a cell type. We began with an examination of the well-studied *Drosophila* anterior-posterior (AP) segmentation system. We trained an activity model using 46 characterized enhancers and 10 TFs known for their role in A/P patterning [50]. With our model, we annotated putative regulatory sequences for other A/P expressed genes and examined the specific edges in the underlying regulatory network. We next attempted to learn a model of enhancer activity for 195 poorly characterized cell types of

Drosophila development [1] for which there is scant knowledge of functional enhancers and relevant TFs. For 77 of these cell types, we were able to construct predictive models from putative enhancers using the expression and motif-based binding of predicted regulatory TFs as well as chromatin accessibility information [23]. Our method to identify a collection of specific regulatory sequences in a novel cell type has the advantage of only requiring a single experimental assay (on chromatin accessibility), as opposed to hundreds of ChIP-seq assays required by its closest alternative.

1.4 Characterizing gene sets with algorithms using heterogeneous biological networks

Typically co-expressed gene sets are examined for enrichment with different "properties" (genes of particular biological process, pathway, or other experimental condition) independently, one at a time. This process ignores the potential relationships between the properties as well as relationships among the genes themselves. The dependencies between the properties and genes may be exploited to reinforce the statistical association between a novel gene set and an annotation that the gene set is enriched for [51], and may be able to better characterize closely related genes. For example, if a novel gene set is enriched for properties P1 and P2, and the orthologous gene set in another species is enriched in the same two properties, then any gene with strong signals for P1 and P2 are likely to be related to the novel gene set. These types of finding may only be possible by combining multiple evidences. Frequently, this combination is achieved by building networks of biological knowledge. There are many network-based approaches for ranking genes for their relationship to a given gene set based on multiple, biological evidences [52]. Some approaches collapse all properties onto a single homogeneous gene-gene network [51, 53, 54]; while others rely on the simple relationships between their evidence types to create a heterogeneous network with two or three edge or node types [55, 56]. GeneMANIA [57] employs an approach that constructs multiple homogeneous gene-gene networks, one for each property type, and given a novel gene set, combines the networks to find the most related genes.

In Chapter 6, we present a method for ranking genes related to a given co-expressed gene set in the context of a large, heterogeneous collection of characterized gene relationships and properties. Our method builds an initial network with multiple node and edge types, preserving more of the original, specific property information than the methods describe above. In the first stage of our novel algorithm, we find the properties that are the most relevant to the co-expressed

gene set. We applied this knowledge to extract a subnetwork of the original network only containing relevant properties. In the second stage, we report the rankings of genes related to the co-expressed gene set based on a random walk with restart on the relevant subnetwork. We demonstrated the effectiveness of this algorithm for ranking genes related to embryonic *Drosophila* development [1] and aggressive responses in the brains of social animals [7].

2 BACKGROUND

2.1 Basics of gene regulation

To understand why genes are expressed in similar patterns, we need to understand the basics of transcriptional regulation. Each cell of an organism contains a copy of the DNA sequence containing the genomic instructions for all of its necessary biological processes. Typically, a small percentage of the genome will encode the instructions for assembling new proteins, which drive development and function within an organism. For example, the ~23,000 genes in the human genome only represent about three percent of the DNA sequence. Transcription is the intermediate process before protein production in which genetic DNA is transcribed in the nucleus into mRNA. To initiate the transcription process, RNA-polymerase enzymes will bind to the "transcription start site" (TSS). In eukaryotes, the mRNA may be spliced with a subset of the protein encoding regions, exons, preserved. The mRNA then undergoes the process of translation into a sequence of amino acids, which will fold into a functional protein.

Protein production from a gene (also called 'gene expression') may vary between cells of different tissues in an organism. Instructions for the control of cell type-specific gene expression are often found in the nearby non-coding DNA regions. These regulatory sequences of DNA interpret the biological condition of the cell to control the timing and quantity of gene expression. In the transcription process, the RNA polymerase's initiation of transcription is affected by the presence of other proteins, known as transcription factors (TFs). The transcription factors bind to short 5-15 base pair sequences called transcription factor binding sites (TFBS) in the neighboring non-coding regions of the gene. When these bound proteins attract the RNA-polymerase and increase the rate of transcription, they are known as activators. When the bound proteins limit transcription by directly or indirectly hindering the ability of the RNA-polymerase to bind to the promoter, they are known as repressors. A transcription factor may bind to similar, but not identical base pair sequences. The DNA binding specificity of a TF, or motif, is often approximated by a position weight matrix (PWM). A *cis*-regulatory module (CRM) or "enhancer" is a homotypic or heterotypic cluster of transcription factor binding sites that act in concert to regulate gene expression [58]. Validated *Drosophila* and mouse enhancers from the REDfly [59] and VISTA [60] databases are around 500 to 3000 bp in length. Heterotypic clusters of binding sites found in complex organisms are the signature of combinatorial regulation of

genes involving interactions of multiple transcription factors. Additionally, eukaryotes may have multiple enhancers per gene, each affecting expression in one or several cellular conditions [61]. While often found near the transcription start site of a gene, enhancer sequences have been shown to affect gene transcription even from distances of several hundred kilobases [32].

A gene expression pattern refers to the spatial or temporal localization within and among cells and tissues where transcription of the gene occurs. An expression pattern may be composed of a single or multiple domains. Cells or tissues where genetic transcription is not occurring are called non-expressed regions of a gene. Regulation of gene expression patterns depends on both the presence of transcription factor binding sites in an enhancer and the presence or absence of the transcription factor proteins in different cells or tissues. The enhancers of a gene may independently drive separate domains of expression because their TF inputs have non-uniform concentrations and the modules are composed of dissimilar configurations of TFBS. Regulation of a gene's expression is also affected by the three dimensional shape of the DNA sequence and other various proteins bound to it in a cellular condition, also referred to as chromatin structure. When DNA is tightly wrapped around histone protein complexes forming nucleosomes, TF binding and gene transcription are hindered. On the other hand, loosely packed, more "accessible" DNA regions may be bound by regulatory proteins, and nearby genes may be actively transcribed [62]. Transcription factors known as pioneer factors are thought to find inaccessible regions of chromatin, disassemble the nucleosomes, and enable other TFs to bind to their cognitive sites in previously inactive enhancers [63].

2.2 Methods for characterizing components of regulation

There are many important technologies in use today for characterizing the genome in specific cellular contexts. As discussed in Chapter 1, hybridization based microarray [2] and sequencing based RNA-seq [3] technologies quantify the amount of a gene's mRNA that is being transcribed. Fluorescence in situ hybridization [1] assays additionally provide complex spatial patterns of gene expression. Patterns of enhancer driven expression are often identified by incorporating reporter constructs containing regulatory sequences and a gene encoding a fluorescent protein into the genome [60]. There also exist high throughput methods, like STARR-seq [64], that are able to quantify the expression levels driven by millions of candidate enhancers in parallel using barcodes or self-transcribing reporter constructs and sequencing.

There are also several experimental assay for assessing the state of chromatin within a cell. In the DNase-seq method [65], accessible regions of the genome that are available for enzymatic cleavage are isolated and sequenced. When sequenced to a great depth, these experimental methods are even able to show individual transcription factor binding sites that were protected from cleavage by bound TFs [66]. Formaldehyde-assisted identification of regulatory elements followed by deep sequencing (FAIRE-seq) is an alternative method for discovering accessible regions by sequencing regions that were not bound to cross-linked proteins [67]. Chromatin accessibility is often predictive of enhancers in a particular cell type [23]; however, it has been noted that some accessible regions are sites of insulator proteins and other transcriptionally repressive elements [64].

Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) is the standard method for measuring the level of TF binding (or "occupancy") at each position in the genome [68] . In a ChIP-seq experiment, DNA bound to a TF is isolated and sequenced. Newer methods, like ChIP-exo [69], are able to improve the resolution of the sequencing results and more precisely identify the TF binding peaks. ChIP-based experiments show that TF binding occurs throughout the genome, often in common, non-specific regions called high occupancy target (HOT) regions [70, 71]. ChIP technologies are also able to identify sequences bound to histone proteins containing various post-translation modifications. Active enhancers are often found in regions with H3K4me1 and H3K27ac marks. Other histone marks like H3K27me3 typically denote inactive regions of tightly compacted DNA [72].

Computational analysis provides a complementary means to discover functional enhancers in the genome. These methods rely on experimentally characterized DNA-binding specificities for each TF. There are a number of methods to characterize the binding specificities of transcription factors. Protein binding microarrays measure the level of TF binding of the protein to each possible 10-mer DNA sequence to characterize the motif [73]. [74] employs a high throughput SELEX method that involves multiple rounds of isolation and amplification of short sequences bound by the TF. The bacterial-one hybrid strategy [75] creates a system in *E. coli* where only clones containing constructs with TF target sequences will have a survival advantage. Far more TFs have had their motifs characterized with *in vitro* assays than have been subjected to ChIP-seq analysis [76]. For example, while about 60% of the nearly 1400 human TFs have motifs available today [74], less than 10% of human TFs in the ENCODE project [77]

have ChIP data available in a limited number of cell types/lines, though the number is growing. It is reasonable to expect that in the near future, most TFs in human and certain model organisms will have characterized motifs either from direct experimental assay or by imputation via homology. Initial work [78, 79] demonstrates the possibility of using these motif collections to perform regulatory analysis on less studied organisms. Later chapters will apply computational techniques to take these characterized TF motif specificities and predict TF binding and annotate the activity of cell type specific enhancers.

Finally, experimental assays of three-dimensional spatial proximity in chromatin regions improve mapping of enhancers to their regulatory gene targets. First, ChIP methods identify binding of insulator proteins that induce chromatin looping. Enhancers have been shown to regulate genes within the same chromatin loop but not genes outside of the insulator boundaries [80]. Additionally, chromosome conformation capture methods (3C, 4C, 5C, Hi-C [81]) identify regions of the genome that are in close physical contact. In these experimental assays, proximal regions are cross-linked and are identified through sequencing. These methods help identify which gene is regulated by an enhancer in a cellular condition. However, the chromosome conformation capture methods currently have relatively poor resolution and are difficult and expensive to apply at the genome-wide scale.

2.3 *Drosophila* embryonic development

In Chapters 3, 5, and 6, we will apply our methods to the well-studied system of *Drosophila* embryonic development. The first fifteen hours of embryonic development after fertilization are divided into 16 stages. In the first three stages, the embryo undergoes nine rounds of nuclear division. In stages 4-6, the blastoderm (a single cell with hundreds to thousands of nuclei) undergoes cellularization and cellular membranes between the nuclei form. Gastrulation of the blastoderm forms the mesoderm, endoderm, and epidermis during stages 7-8. Features of the head begin to develop in stages 9-10. In stages 11-12, the anterior and posterior midgut fuse and body segments are initially observable. In the final studied stages (13-16), the central nervous system and most other organ primordia differentiate [82]. Images of *in situ* hybridization of over 7,000 genes in each of these developmental stages were cataloged by the Berkeley *Drosophila* Genome Project [1]. Genes are annotated by the differentiated organs in which they are expressed (e.g., malpighian tubules). If the organ has not yet differentiated, but has a distinguishable morphology, the gene is annotated as being expressed in the organ

primordium (e.g., malpighian tubules primordium). Finally, if the gene is expressed in nuclei or cells that are morphologically indistinct but will eventually give rise to a particular organ, the gene is annotated as *anlage in statu nascendi* (e.g., dorsal ectoderm ASN). Over 195 different developmental stage and annotation term combinations were assigned to thousands of genes in this expression database.

A particularly well studied system within the *Drosophila* blastoderm is anterior/posterior segmentation. This system has well-characterized transcription factors that act in a hierarchical structure to generate increasingly complex gene expression patterns along the A/P axis of the embryo, which eventually results in the segmented body plan of *Drosophila* adults. The examined transcription factors in this pathway are classified into three groups: maternal, gap, and pair-rule. These categories loosely capture the temporal development of the network, with the transcription factors encoded by the genes in earlier groups being a prerequisite for the expression of the later genes. The mRNAs of the maternal genes are deposited in the oocyte before fertilization. For example, the mRNA of the maternal gene *bicoid* (*bcd*) is localized during oogenesis in the future anterior of the embryo. In the zygote after translation, the BCD protein will be present in a decreasing anterior to posterior concentration gradient. The overlapping combinations of concentration profiles (expression patterns) of the maternal factors will activate regions of expression of gap genes along the anterior/posterior axis. Along with maternal inputs, the domain boundaries of later gap gene expression are partially regulated through other gap factors. The last group of genes to show expression before the formation of cellular membranes is the pair-rule genes. "Primary" pair-rule genes are initially regulated by maternal and gap transcription factors. The expression of the pair-rule genes is typically expressed in seven anterior/posterior domains across the embryo. Primary pair-rule genes also serve to regulate the later "secondary" ones. Our focus on the known *cis*-regulatory modules of the segmentation network forms the basis of the study in Section 5.2.

3 PREDICTING REGULATORS OF EXPERIMENTAL GENE SETS

This chapter introduces a novel pipeline for identifying potential transcriptional regulators of co-expressed gene sets. The pipeline is described in greatest detail in [83] from the 2014 Web Server edition of Nucleic Acids Research. The procedure for incorporating sequence conservation in Section 3.2.1 was part of a joint work with Majid Kazemian that was published in PLoS Biology [84].

3.1 Background

Few tools exist that take an experimentally derived gene set and examine their corresponding non-coding regions for evidence of a shared regulatory signature. This is an important analysis that uncovers major transcriptional regulators of the novel gene set and suggests mechanistic explanations for the results of the experiment. Some of these tools [15, 22, 85] are designed to identify major regulators of novel gene sets using data from experimental assays, especially ChIP-seq data. The problem with relying on these approaches is that generating ChIP-seq data to exhaustively identify all gene set regulators in a cell type is too time consuming and expensive to be feasible. For example, the well funded ENCODE [86] and ModENCODE [17] consortiums only produced ChIP-seq data for tens of transcription factors in a limited number of tissues. There are also technical issues such as the amount of sample required or characterizing efficient antibodies that make producing ChIP-seq data difficult for many tissues and organisms of interest.

De novo motif discovery is another common approach to identify relevant motifs from the regulatory sequences of a novel gene set. There are several tools that implement this type of search for overrepresented sequence patterns [18, 19, 87]. The problem with these methods [20] is that the significance of the results is lessened by the large space of solutions searched and that the biological interpretation of the results is often difficult.

In this chapter, we present our pipeline for the regulatory signal enrichment task, which assembles the sets of genes that are likely to be regulated by each transcription factor and quantifies the significance of their overlap with the experimental gene set. Our method relies on computational prediction of transcription factor binding based on DNA sequence and characterized TF DNA-binding motifs. This allowed us to search for the signature of the hundreds of transcription factors whose motifs have been experimentally characterized. Computational prediction of TF binding is susceptible to high false positives, especially in large

genomes. Our approach is able to mitigate this problem by incorporating additional data such as binding site clustering and sequence conservation. Our method is also sensitive to the G/C content of the regulatory sequence and the transcription factor motifs, which was demonstrated to be important in [88].

We found, like [43, 89], that chromatin accessibility data from the related cell type significantly improves the accuracy of our computational predictions. This experimental data must only be generated once for each cell type, not for each transcription factor as with ChIP assays. We go beyond studies [23, 43, 89-95] that explore how well motifs and/or accessibility data predicts ChIP-based occupancy profiles to assess how these approaches fare in the ultimate goal of identifying relevant TFs. These evaluations were primarily done in the well-studied system of *Drosophila* embryonic development. Ultimately, we found that our method to identify transcriptional regulators of novel gene sets compares favorably to methods that rely on ChIP-seq data and was able to identify regulatory characteristics of TFs and co-expressed genes. We applied our method to several other systems and made an online web tool Motif Enrichment Tool (MET) available to researchers.

3.2 Computational prediction of TF binding

The first step in our method to identify transcriptional regulators of an experimental gene set is to produce computational predictions of genome-wide TF binding profiles. We begin by masking the tandem repeats in the genome of interest with the Tandem Repeat Finder [96]. Tandem repeats are short, repetitive DNA sequences non-uniformly interspersed throughout the genome and not known to be important in transcriptional regulation. For this reason, repeat masking has the effect of ignoring regions where the experimentally characterized DNA-binding specificities of TFs ("motifs") may match the pattern of the tandem repeats.

We next take a TF motif from one of several public databases [73, 74, 97-99] and create a genome-wide scoring profile of that TF's binding using the motif and computational motif scoring software. Our profile assigns a score to every 500 bp window in the genome (in shifts of 50 or 250 bp depending on genome size), representing the strength of that motif in the window. These 500 bp windows represent potential enhancers, the major regulatory sequences embedded in the genome. In the REDfly [59] and VISTA [60] enhancer databases, the common and minimum size of characterized enhancers is approximately 500 bp. It is also the size adopted by

enhancer finding tools like PhylCRM [100] and experimental enhancer finding techniques like STARR-seq [64].

To increase TF binding efficiency, enhancers are likely to contain homotypic clusters of binding sites. For this reason, we score each genomic window for a motif with the HMM-based program Stubb [101]. Stubb computes a single score that integrates over all strong and weak matches to the motif present in a window. We typically run Stubb with a fixed motif state transition probability of 0.0025 and with a set of 5 kbp upstream or gene desert sequences from the genome to train the background model. Once we have scored every window in the genome for the motif, we record the average and standard deviation of the genome-wide Stubb scores for the particular motif.

3.2.1 Incorporating sequence conservation

Another important assumption of enhancers is that in order to maintain their function across species, they will conserve their TF binding sites/content at higher rates than non-functional regions. Following this assumption, we designed a novel method for phylogenetically averaging the motif scores computed from orthologous genomic regions. In principle, this method enables us to remove false positive high scoring windows because only true enhancers will have a high score across multiple related species. Our approach models the motif score of a region as a random variable evolving through Brownian Motion dynamics [102] along the branches of the phylogenetic tree and computes the expected tree-wide average of this variable given its observed values in the extant species. The computation of this "Brownian Motion average" required a novel implementation of the "upward-downward" algorithm [103].

Our novel procedure takes a phylogeny, *T*, and a motif score value for each extant species at the leaf nodes. We describe our method with the following notation:

- X_i : random variable representing trait (observation) at node i
- $\pi(i)$: parent of node i
- $T(i)$: subtree rooted at node i
- t_i: branch length between node i and its immediate parent
- O_1: all observed traits in the tree
- $O_i = \{x_n | n \in T(i) \text{ and } C(n) = 0\}$: all observed traits in subtree rooted at node i
- $O_{i\setminus j} = \{x_n | n \in T(i) \text{ and } n \notin T(j) \text{ and } C(n) = 0\}$: all observed traits in subtree rooted at node i but not at node j

In order to calculate the "Brownian motion average", we need to take a temporal average of the random variable *X* over the entire phylogenetic tree. We calculate this as the sum of the expected values of each branch, $E_{i,\pi(i)}$, weighted by its branch length, t_i :

$$\frac{1}{2}\sum_{i \in T(1)} t_i E_{i,\pi(i)}$$

To define the expected value of the branch, $E_{i,\pi(i)}$, we average the expected value at each of its endpoints, *i* and *π(i)*, of the random variable *X* given all of the observed values at the leaves.

$$E_{i,\pi(i)} = \frac{1}{2}[E(X_i|O_1) + E(X_{\pi(i)}|O_1)]$$

We utilize the upward-downward algorithm and our assumption of Brownian Motion to compute the expected value of the random variable at any node, $E(X_i|O_1)$. The upward-downward algorithm produces two probability distributions for every node on the tree. The first is the "downward" probability distribution $\alpha_i(m)$ that captures the probability of the random variable taking the value *m* at node *i* and the observations at the leaves not under node *i*.

$$\alpha_i(m) = \Pr(X_i = m, O_{1\setminus i})$$

This "downward" probability has a recursive formulation that is computable given its values higher in the tree. The "upward" probability distribution $\beta_i(m)$ captures the probability of the observations at the leaves under node *i* given the value of the random variable is *m* at node *i*.

$$\beta_i(m) = \Pr(O_i | X_i = m)$$

It has a recursive formulation that depends on its values lower in the tree. The product of these two distributions at a node *i* is the joint probability of all observations and the value of the random variable at node *i*.

$$\Pr(X_i = m, O_1) = \alpha_i(m)\beta_i(m)$$

Since we model the evolution of our random variable with Brownian Motion, we will always be able to represent its probability distribution with a Gaussian probability distribution function. This relies on the important identity that the product of two Gaussian distributions is a Gaussian. In our framework, the calculation of probability distributions $\alpha_i(m)$ and $\beta_i(m)$ are Gaussian, so the joint probability distribution $P(X_i = m, O_1)$ is also a Gaussian represented generically with the notation $N(m;\ \mu_i, \sigma_i^2)$. The expected values of the random variable at any node, $E(X_i|O_1)$, we use to calculate our phylogenetic averages are just the μ_i means of the joint Gaussian distributions returned by the upward-downward algorithm. The recursive definitions of the upward and downward probabilities as well as a more explicit derivation of the result are found in the supplementary materials of [84].

In practice, we create a Brownian Motion average motif profile by first performing the Stubb scan independently in each species and converting the Stubb score profiles into z-score profiles by subtracting the corresponding genomic average and dividing by the standard deviation. We then map the scores from all auxiliary species to the coordinates of the genome of the species of interest. For every window in the genome of the species of interest, we take the phylogenetic tree and the non-negative z-scores of the window and its orthologs to compute the corresponding Brownian Motion average. Our multi-species motif profiles in *Drosophila melanogaster* are computed from scores of genomes for 11 species of flies [104].

3.2.2 Normalization of motif scoring profiles

The next step in our pipeline of producing computationally predicted profiles of TF binding is to rank-normalize the single or multi-species scoring profiles, converting the original motif scores into scores from 0 to 1 where 0 represents the best value. This is helpful for comparisons across multiple motifs. The range of Stubb scores for two different motifs may vary significantly depending on the complexity of the motif. However, the rank normalized window score of 0.01 means that the window is in the top 1% genome-wide for that motif, regardless of its complexity. We also perform at this stage a variant of this normalization procedure, which considers the local G/C content. The motivation is straightforward. If a motif is composed of

mostly C's and G's, then a high Stubb score is expected to be computed in a G/C rich window. We are interested in those windows where the motif matches are much stronger than expected by G/C content alone. Thus, the 'G/C normalization' procedure separates genomic windows into 20 equal-sized bins based on their G/C content and performs rank-normalization within each bin separately. In the study of honeybee behavioral genes [88], a significant G/C bias was discovered in the promoters of the genes. Only after applying the G/C normalization procedure were a majority of the spurious results from this confounding signal eliminated.

3.2.3 Chromatin accessibility filters

Another method for increasing the accuracy of predicted TF-DNA binding profiles is using cell type specific chromatin accessibility data as a filter. It combines the static sequence-encoded information about TF-binding potential of motif scoring with the dynamic, tissue or stage-specific data from chromatin accessibility. Chromatin accessibility is characterized with the DNaseI hypersensitivity [23], FAIRE-seq [67], or ATAC-seq technology [105]. It may also be inferred from ChIP-seq characterized histone modifications and other epigenetic marks (reviewed in [32]). Chromatin accessibility information for several tissues from fruit flies and humans are available from the BDTNP [106] and ENCODE [86] projects. We download raw chromatin accessibility data for a specific tissue and create an average accessibility score for each 500 bp window in the genome. While the percentage of the genome that is accessible and functional may vary across cell types and species, we rely on estimates from the developing *Drosophila* embryo [23] and consider only motif scores that fall within the top 10% of accessibility as potential enhancers in that tissue. Practically, this means all windows not within the top 10% of accessibility scores have their Stubb score re-assigned to 0. We call the scoring profiles that have been filtered by this chromatin structure data our "motif + accessibility" scores.

3.3 Evaluation of computational TF profiles

The current gold standard for predicting regulatory roles in gene expression is TF occupancy data from ChIP-seq experiments. We determined to show that these data could be substituted with computational TF motif scans, especially when complemented with cell type specific chromatin accessibility data. We began with 69 ChIP datasets covering over 40 TFs in various stages of *Drosophila* embryonic development [17, 38, 107-110]. The raw ChIP data was converted into averaged values for each of our 500 bp genomic windows. For each ChIP dataset,

we selected 1000 non-overlapping ChIP peak genomic windows and 1000 random, non-coding windows. Consistent with previous studies [17, 107], we found that most pairs of TFs have very highly correlated binding profiles. This is commonly attributed to the strong influence of chromatin accessibility on TF binding [23].

In order to create computational motif scoring profiles, we started with a collection of DNA binding specificities characterized with the bacterial one hybrid (B1H) technology made available by FlyFactorSurvey [99]. This collection contained 325 motifs of distinct fly TFs. We produced single species scores for every motif as well as multi-species scores from the Brownian Motion averages on 12 *Drosophila* species. We also downloaded DNaseI-seq chromatin accessibility data from BDTNP [106] from five stages of fly embryonic development (5, 9, 10, 11, and 14) to serve as stage-specific chromatin filters for our computational motif scores. We examined the correlation between the ChIP scores of the 2000 windows of each dataset and the corresponding single species "motif + accessibility" scores, and found an average Pearson correlation coefficient (PCC) of 0.52 across the 69 datasets. This average correlation improved to 0.66 when we incorporated multi-species "motif + accessibility" scores, with 61 of the 69 datasets having a PCC > 0.5 (Figure 3.1). This is an intriguing observation since the ChIP data reflects binding specific to *D. melanogaster*, however, we speculate that evolutionary conservation serves as a proxy for the contextual information that is necessary for *in vivo* TF binding.

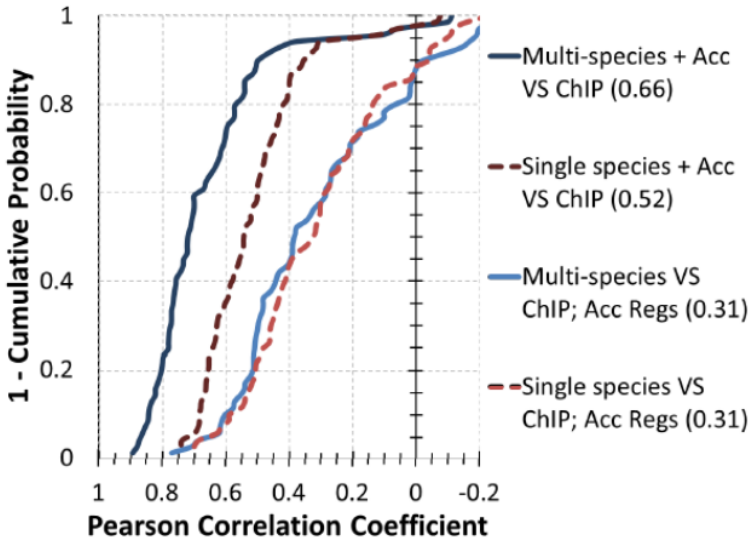


Figure 3.1 Evaluations on 69 ChIP Datasets. Each line plots for a given correlation value (x-axis) the percentage of the 69 ChIP sets (y-axis) that are greater than that correlation value. The evaluations using multi-species (single-species) scores are solid blue (dotted red) lines. The darker lines represent evaluations between ChIP scores and

"motif + accessibility" scores, while the lighter lines represent evaluations comparing ChIP scores to "motif only" scores in only accessible regions.

We wanted to separate the improvement in the correlation due to accessibility from the specific TF motif that produced the scores. To do this, we generated a second set of 2000 windows for each ChIP dataset, this time additionally requiring that each window be in the top 10% of accessibility for the matching developmental time point. We observed an average PCC of 0.311 for multi-species motif scores with ChIP scores in accessible regions only. Forty of the 69 datasets had a PCC greater than 0.3, confirming that motifs are highly informative of TF-DNA binding levels, even within accessible regions of DNA. We also noted negative PCC values in 8 of the 69 datasets similar to some previous reports [17, 89]. Many of these negative instances occurred with ModENCODE ChIP datasets, which may in part be because these datasets often correspond to relatively broad developmental intervals and in part due to technical limitations in some of these assays. We tested the multi-species motif scores for specificity to the appropriate "accessible regions only" ChIP dataset. Figure 3.2 shows that in most cases the score predictions from the corresponding motif exhibits greater concordance with its ChIP dataset than the predictions from motifs of different TFs. Our results support the premise that TF motifs together with accessibility data may approximate TF-DNA binding profiles in instances where ChIP assays on multiple TFs may be impractical.

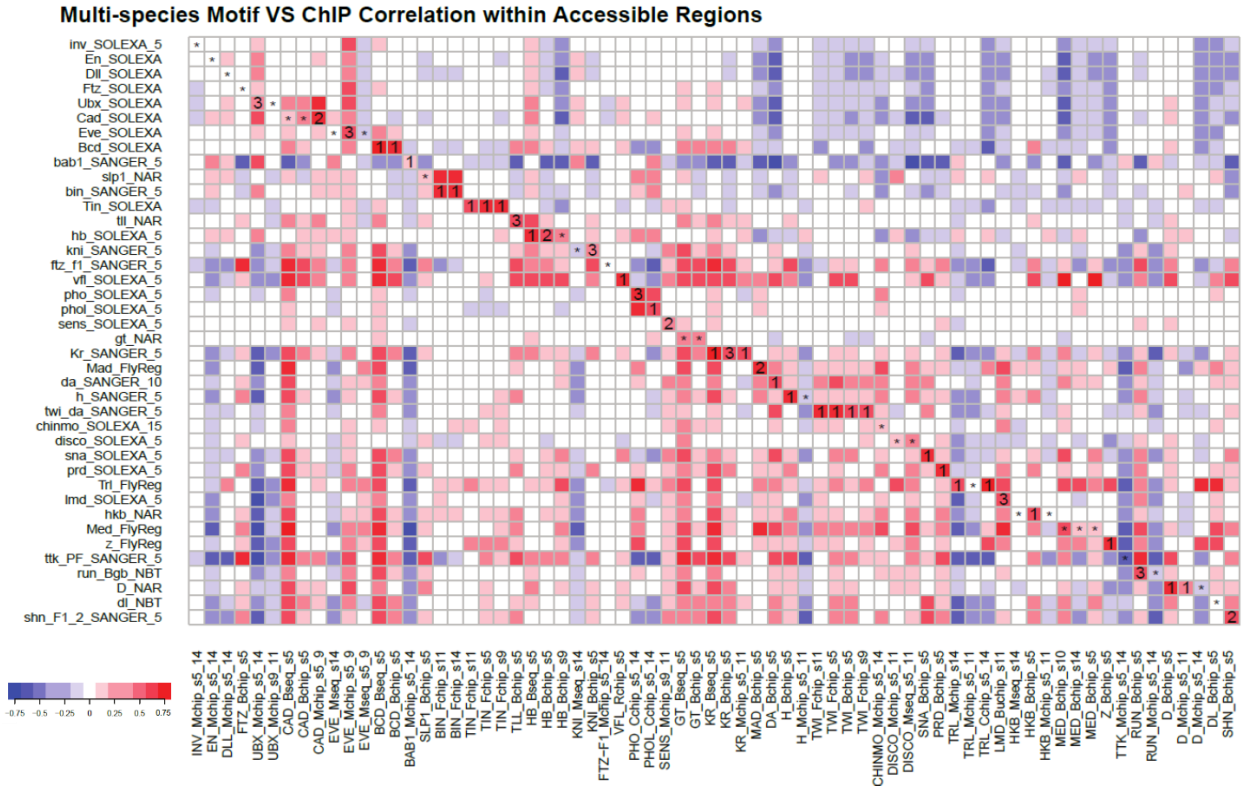


Figure 3.2 Correlation between Motif and ChIP Scores in Accessible Regions. The columns of the heatmap represent the 69 ChIP named for the assayed TF, laboratory source, and developmental stage. The rows represent the experimentally determined motifs of the 40 corresponding TFs. Each cell is colored for the Pearson correlation between 2000 accessible windows selected to have 1000 non-coding ChIP profile peaks and 1000 non-coding random regions. In a cell where the motif and ChIP profile represent the same TF, the rank (or star if rank > 3) of that motif by its correlation among the 40 TFs is enumerated.

We examined if any multi-species motif scores correlate with chromatin accessibility scores alone as this might be anticipated for pioneer factors that establish a permissive chromatin state [63]. For each DNaseI-seq dataset from a distinct developmental stage, we selected 1000 non-overlapping accessibility peak genomic windows and 1000 random, non-coding windows. We found several motifs with strong positive correlation to accessibility scores; including known pioneer factors such as *Trithorax-like* (TRL) [111] and *Vielfaltig* (VFL) [112], also called *Zelda*, as well as basic helix-loop-helix TFs such as *Medea* (MED), and *Mothers against dpp* (MAD). Surprisingly, many of these correlations are comparable to or even better than the correlations between the motif-based scores and their corresponding ChIP profiles. We observed clear trends in time-dependent roles of motifs in predicting accessibility, e.g., VFL is correlated primarily at the earliest stages of development and TRL increases in importance during later stages, as has also been reported previously [89, 112]. Interestingly, there were also several homeodomain TFs, including *Bicoid* (BCD), *Caudal* (CAD), *Engrailed* (EN), and *Invected* (INV) that are negatively

correlated (SPCC ≤ -0.35 over 2000 windows, p-value ≤ 1E-56) with chromatin accessibility, a phenomenon for which we are unaware of any suggested mechanisms in the literature. Overall, our analysis of accessibility data strongly suggests the potential of a motif-based computational method to approximate accessibility profiles, as long as the relevant motifs are identified for the cell type of interest.

3.4 TF target set construction and enrichment tests

The main goal of this chapter is to identify major regulators of novel gene sets. We approach this problem by computationally predicting the sets of genes that are likely to be regulated by each transcription factor and quantifying the significance of their overlap with the gene set of interest. In the proceeding sections, we demonstrated an accurate method for the prediction of TF binding genome-wide using the TF's characterized DNA-binding motif and available cell type specific chromatin accessibility data. The rest of this chapter will focus on how to define the set of genes targeted by each TF.

To define the "target gene set" of the TF, we identify the genes that have the strongest profile scores in their regulatory regions. When TF motifs are relied on to generate the TF binding profiles, we call the identified gene sets "motif target sets" or motif modules. Our procedure also is able to process the windowed ChIP occupancy profiles in order to make ChIP-based TF target sets. The most important parameter in defining these gene sets is the definition of the regulatory region. The most common regulatory region definitions involve predefined, fixed lengths around the transcription start site (TSS), e.g., "1 kbp upstream", "10 kbp upstream", or "5 kbp upstream and 2 kbp downstream". We also create several definitions of regulatory regions of variable length. The "nearest TSS" defines regulatory regions as the genomic windows that are closer to the gene's TSS than to any other TSS. This maps every window in the genome to a single gene, with closely packed genes only receiving a few windows apiece. The "gene territory" regulatory region definition includes the gene's body and half the distance to the nearest non-overlapping genes upstream and downstream, with a minimum of 5 kbp included upstream. This enables each gene to map to several windows, but it also means that windows are able to map to more than one gene. We also allow for regulatory region definitions based on experimental assays of the chromatin. For example, in *Drosophila*, we define the "intergenic" (IG) control region suggested in [80]. We downloaded the 1% FDR ChIP-chip data for three known insulator proteins (BEAF-32, CP190, and CTCF_C) for early embryonic

development (E0-12h) from ModENCODE [17]. For each gene, the IG regulatory region includes the gene and extends on either side of the gene by 50 kbp or until a window in which two of the three insulator proteins are bound, whichever happens first [80]. Short regulatory region definitions are likely to contain most enhancers in compact genomes. Longer regulatory region definitions (>5 kbp) are more likely to be noisy, but necessary to capture distal enhancers in large vertebrate genomes. The noise introduced by large regulatory region definitions increase the importance of incorporating additional motif prediction filters such as conservation and chromatin accessibility.

Once a regulatory region type has been selected, we produce a score S_g for each gene g for the presence of a given *cis*-regulatory feature in that gene's regulatory region. This is given by:

$$S_g = 1 - (1 - P_g)^{w_g}$$

where P_g is the best normalized score (either among all or windows of similar G/C content) of the regulatory feature in the regulatory region of g, and w_g is the number of windows in the region. The best normalized window score P_g is an empirical p-value that will be between 0 and 1, with an approximately uniform distribution, and may be interpreted as the probability that a random window will score as well or better for the given motif. When one interprets P_g this way, S_g is just the probability of finding the minimum p-value of P_g when given w_g IID p-values. Figure 3.3 shows the normalized score profile for a single motif in the regulatory region of gene g filtered by chromatin accessibility and the components for computing the corresponding regulatory feature score, S_g.

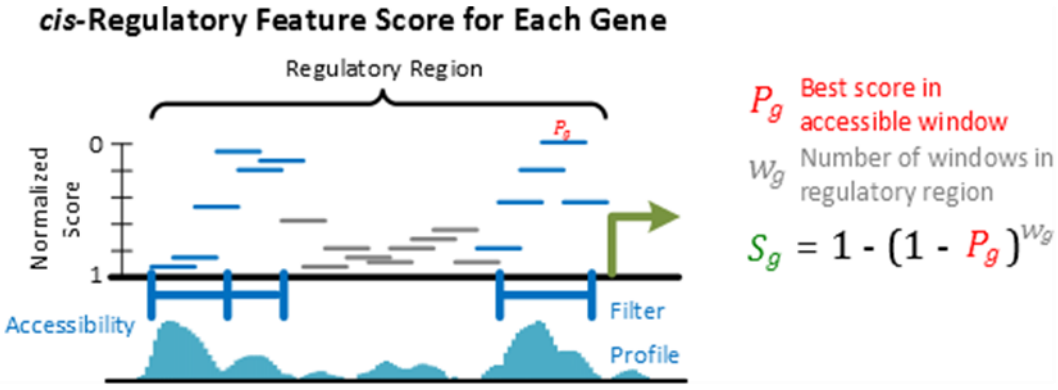


Figure 3.3 Calculation of Motif Score for Gene. Shown is an upstream regulatory region of a gene g. The horizontal bars within the regulatory region represent the normalized motif score of each window for a single TF motif with higher bars representing better scores. Below in blue is a chromatin accessibility profile of the genomic locus and the thick blue bars indicate accessible regions of chromatin in which we will consider motif-based scores.

The regulatory feature score of g for this example TF motif is calculated as shown on the right using the best score, P_g, in the accessible windows of the regulatory region.

To create the final target gene set corresponding to the feature (motif or TF), we select a fixed number of genes (often 400) with the best S_g scores. We do not claim that these 400 genes are in fact the direct regulatory targets of the TF, or that every TF has the same number of targets. Rather this methodological choice is made in order to ensure parity among the many enrichment tests (one for each TF). We have tried different thresholds on the number of genes or the values of S_g, but these did not improve our evaluations.

After defining the target gene set of the motif, we quantify the significance of its overlap with the original gene set of interest using the p-value of the one-sided Fisher's exact test. This is the standard approach using the hypergeometric distribution employed by ~60% of the 68 enrichment tools surveyed in [9]. The motif target sets are constructed and tested for every motif in the collection with the top results assigned as the major regulators of the co-expressed gene set predicted by our method. Evaluation of the results may be necessary to ensure the proper selection of the parameters of the algorithm (e.g. the choice of normalization procedure, the definition of the regulatory region, or the size of the motif target sets).

3.5 Application in fruit fly embryonic development

We next sought to identify the best strategy for discovering TFs associated with a co-expressed gene set and to compare our methods to similar ones using ChIP data. *Drosophila* embryonic development offers an ideal system to evaluate our method because of the relatively mature status of the data types involved; gene expression, chromatin accessibility, TF motif specificities, and ChIP binding profiles. Our first step was to construct 195 co-expressed gene sets from the Berkeley *Drosophila* Genome Project (BDGP) [113]. They have annotated *in situ* images for over 7000 genes in developing embryos for specific "expression domains", tissue or cell types and developmental stage describing the gene's expression pattern. These domains span four developmental stages labelled "4-6", "9-10", "11-12", and "13-16". We only focused on the 195 expression sets that contained between 20 and 1,500 genes and which had a spatially descriptive annotation.

Using our collection of 325 TF motifs and our stage specific chromatin accessibility data, our goal is to identify when a TF plays a broad role in regulating the genes of an expression domain. We call such a statistical finding a "TF - domain association". To test for an association

between a particular TF and expression domain, we started by creating the single or multi-species motif scoring profile from the given TF's motif (see Section 3.3). We then filtered this profile by the DNaseI-seq accessibility data from the stage that corresponds to our expression domain of interest. From this filtered profile, we constructed a motif target set for each one of three regulatory region definitions. The regulatory regions defined in this study are 1 kbp upstream ("p1K") or 5 kbp upstream ("p5K") of the transcription start site, or the insulator-defined "intergenic" regulatory region defined previously ("IG"). The p-values of enrichment were calculated for the overlap of each of the three motif target sets with the expression domain gene set, and the result from the most significant test was recorded as the significance of the "TF - domain association".

To evaluate our TF-domain association pipeline, we collected 3,412 (TF, domain) pairs as a proxy for the ground truth where the TF gene is specifically expressed in the domain. We then evaluated our pipeline by comparing its (TF, domain) pair predictions to the ground truth and reporting the area under receiver operator curve (AUROC). The overview summary of this entire pipeline is illustrated in Figure 3.4.

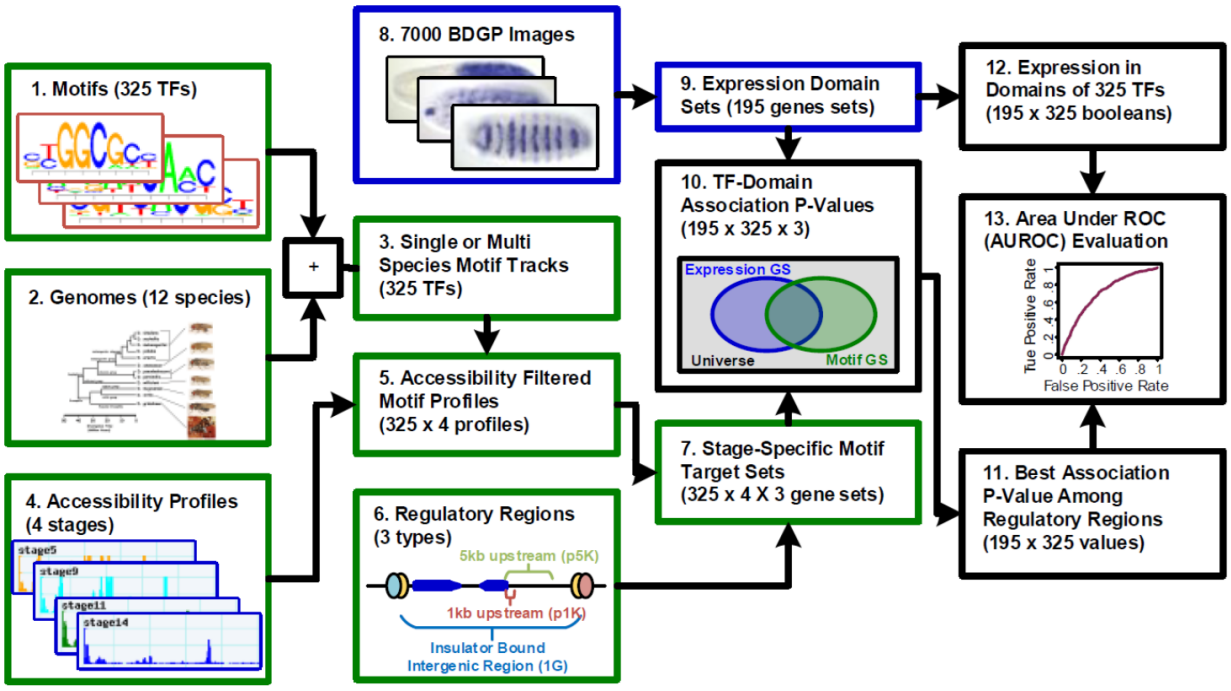


Figure 3.4 Summary of Association Pipeline The association tests are performed between 195 gene sets defined by BDGP expression annotations and gene sets formed from motif scans of 325 transcription factor motifs filtered by chromatin accessibility from 4 developmental stages with 3 different regulatory region definitions. The associations are evaluated by the expression of the transcription factors in the expression domains.

Our results showed that our pipeline using multi-species "motif + accessibility" scores (AUROC = 0.67) was (a) slightly better than when using motif scores from *D. melanogaster* only (AUROC = 0.66), and (b) significantly better than when ignoring accessibility information (AUROC = 0.605). Our strategy of opportunistically taking the best of three regulatory region definitions (p1K, p5K, IG) was found to be slightly superior to any method that only considers one definition alone. At a p-value threshold of 1E-7 (Bonferroni corrected p-value < 0.0064), 5,716 (TF, expression domain) pairs were designated as significantly associated, with a true positive rate of 24% and a false positive rate of 8% based on TF presence in that domain. We also examined how predicted associations based on multi-species "motif + accessibility" scores compare to similar associations that are inferred when we incorporate ChIP scores in their place. We analyzed ChIP datasets from early embryonic development that span 35 distinct TFs, and predicted TF-domain associations among all possible 35 x 195 = 6,825 pairs, using the same approach association pipeline. Using TF expression annotations as ground truth, we were surprised to find that the AUROC of ChIP-based predictions (0.698) was comparable to the motif-based method (AUROC = 0.704, Figure 3.5), all other aspects of the evaluation being the same. We noted the ChIP-based method to have increased sensitivity at high levels of specificity, while the motif-based method recovered more true TF-domain relationships at a 50% false positive rate. The TF-domain associations predicted by these two approaches overlap significantly, with 53% of the 567 ChIP-based associations being recovered from 710 motif-based associations (p-value < 1E-162). This analysis suggests that motif-based approximations of TF-DNA binding profiles are not only strongly similar to ChIP-based profiles, but also that they may be as useful as ChIP data for assigning TFs their regulatory roles in specific expression domains.

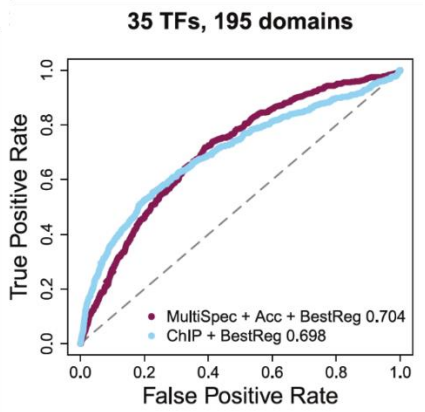


Figure 3.5 Comparing Motif and ChIP Associations. For 35 TFs and 195 expression domains, we compare our "MultiSpec + Acc + BestReg" method of calculating TF-domain associations using the enrichment with the best of the motif target sets defined from three regulatory regions and multi-species motif scores filtered by chromatin accessibility to an equivalent method that instead incorporates ChIP scores. The ROC curves are calculated using domain specific expression of the TF as the ground truth and the AUROC is reported in the legend.

We next focused on the 819 significant TF-domain associations (identified above) that were supported by TF expression data. However, we noted that for a predicted TF-domain association to be concordant with TF expression data, the TF gene need not be annotated with that expression domain. For instance, TFs that are ubiquitously expressed may have a regulatory effect on any expression domain in the corresponding stage. Alternatively, repressive TFs are expected to be expressed in spatio-temporal domains bordering the expression domain of their target genes rather than overlapping it. To discover related expression domain pairs, we downloaded the controlled-vocabulary anatomical term hierarchy from FlyBase [114] and mapped expression domains onto it. We identified 1068 pairs of "related" expressions domains as any two expression domains connected by a relationship type in the term hierarchy with distance one or two. When we considered expression support of the TF in the specific domain, in one of its related domains, and by ubiquitous TF expression in the corresponding developmental stage, we found that 1,232 (22%) of all significant TF-domain associations were supported. Overall, these supported TF-domain associations involved 251 of the 325 TFs that we analyzed and 110 of the 195 expression domains analyzed.

Stage 4-6	Stage 9-10	Stage 11-12	Stage 13-16
TRL (22)	TRL (12)	TRL (24)	TRL (20)
VFL (21)	Z (12)	Z (21)	ADF1 (19)
ADF1 (21)	CG13897 (11)	CG13897 (17)	Z (15)
MED (20)	VFL (11)	ADF1 (15)	DEAF1 (14)
Z (20)	MED (10)	MED (14)	BLIMP-1 (11)

Table 3.1 Commonly Identified Regulators. For each developmental stage, the regulators that are expressed in and significantly associated with the most number of expression domains (in parenthesis) are listed.

The TFs with the most supported domain associations included known pioneer factors VFL and TRL. We also identified *Zeste* (Z) and *Adh transcription factor 1* (ADF1) as important regulators of many expression domains in multiple developmental stages (Table 3.1); both TFs have been linked to regulating polycomb group complexes by binding to polycomb response elements throughout the genome [115, 116]. Many TF-domain associations, such as *Brinker* (BRK) regulating embryonic ventral epidermis, *Twin of eyeless* (TOY) regulating embryonic brain, and *Serpent* (SRP) regulating embryonic/larval fat body, were also corroborated through phenotypic data of mutant alleles curated by FlyBase. As an example, Figure 3.6 illustrates a subset of significant, expression supported TF-domain associations related to the development of the larval feeding organ, clypeolabrum. This regulatory network shows transcription factors that are predicted to be related to all developmental stages (TRL, ADF1), primarily early stages (e.g., *Adult enhancer factor 1* (AEF1), *Sister of odd and bowl* (SOB), VFL), or only later stages (*Tinman* (TIN)), based on motif analysis as well as expression data. The full set of TF-domain associations is made available through an easy-to-navigate online interface at [http://veda.cs.uiuc.edu/B1H_GRN]

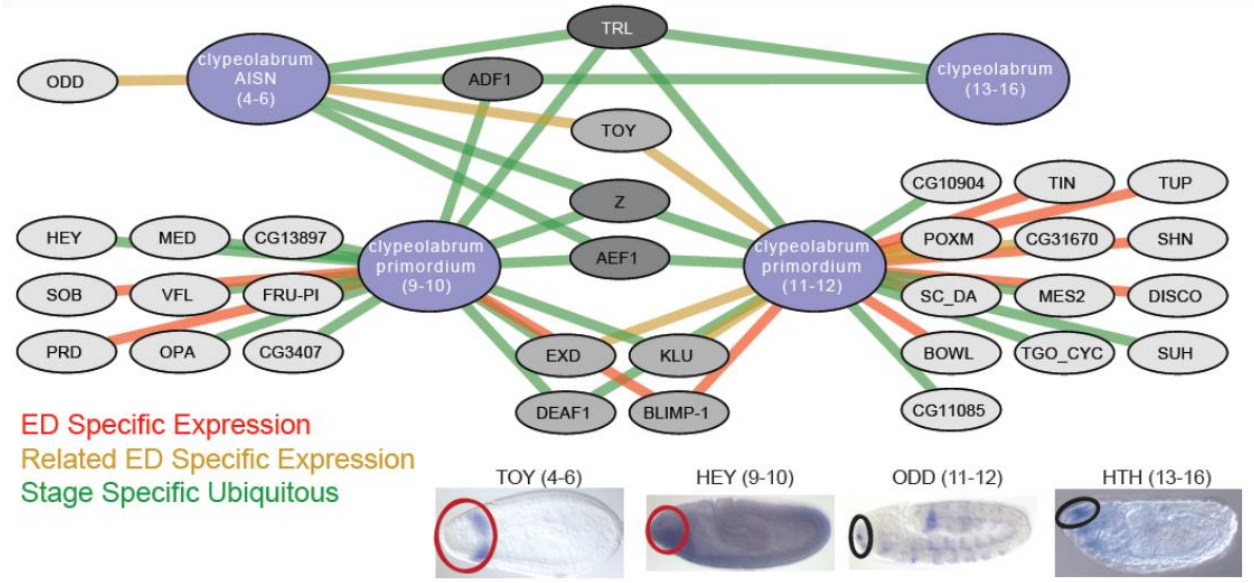


Figure 3.6 Clypeolabrum Network Example. Four expression gene sets from BDGP related to clypeolabrum development in the early embryo are shown as blue nodes ordered counterclockwise from the top left. Grey nodes indicate TFs. Edges are drawn when the corresponding TF-domain association is significant (<1E-7). TF nodes are colored from light to dark by the number of association edges they have. Edges are colored by the type of expression support indicated in the legend. Below the network are *in situ* images of four different TFs at different stages whose clypeolabrum associations are supported with consistent expression (circled).

3.6 Systematic distance biases for regulatory signals

Analysis of TF-domain associations provided systems-level insights into *cis*-regulatory architecture by specifically revealing TFs and expression domains with systematic biases for regulatory regions that are gene-proximal or distal. Our TF-domain associations were based on the strongest association between the expression domain gene set and motif target sets defined from three regulatory regions definitions – 1 kbp upstream ("p1K"), 5 kbp upstream ("p5K") and intergenic with insulator site boundaries ("IG"). Nearly 56% of all significant associations were derived from the p1K definition, which capture only proximal regulatory signals; while for ~28% of associations, the strongest signal came from the "IG" definition, which is of variable length and frequently captures distal regulatory signals. We attempted to quantify if certain TFs or expression domains tend to have stronger regulatory signals in one of these classes of regulatory regions versus others [21, 117].

For every TF-domain pair, we separately recorded the association p-values of the association test using the "p1K"-based motif target set and the "IG"-based motif target set. We converted these p-values to the corresponding z-scores of the standard Normal distribution. To determine if a TF, *F*, had a bias for regulating via proximal promoters, we first counted the number of its expression domain associations out of 195 that were significant (p-value < 0.001) with either regulatory region definition. We define N^{F}_{p1K} as the number of expression domains associated with *F* where the z-score for the "p1K" definition was at least three greater than the z-score of the "IG" definition. N^{F}_{IG} is the corresponding count where the z-score of the "IG" definition is at least three greater. We similarly find N^{ALL}_{p1K} and N^{ALL}_{IG} which count the appropriate associations across all transcription factors. These four numbers define the values of a 2x2 contingency table on which we employ the Hypergeometric test to quantify the significance of N^{F}_{p1K}. A TF's bias for regulating its targets via distal sites was tested in an analogous manner. We examined the expression domains for regulatory biases using the same approaches except counting over the 325 motifs.

Each of the TFs, TRL, *Zeste* (Z), ADF1, *Deformed epidermal autoregulatory factor-1* (DEAF1), CG4360, *Klumpfuss* (KLU), MAD, and MED, were found to have p1K-specific associations, i.e., associations seen only in promoter scans, with over 50 expression domains but no IG-specific associations. *Zeste* has been demonstrated to frequently bind proximally to a gene and facilitate communication with distal enhancers [118]. Alternatively, *Disconnected* (DISCO),

Extradenticle (EXD), *Goosecoid* (GSC), and BCD showed IG-specific associations with tens of expression domains, but few or no p1K-specific associations, thus pointing to dominance of distal regulatory action for these TFs (Figure 3.7). Overall, we found that as a class, homeodomain TFs have a preference for acting via distal regulatory regions, consistent with [21]. We also found several predominantly late-stage expression domains that prefer TF associations with proximal regulatory signals and several early stage domains that are skewed towards distal signals, pointing to an architectural difference between early and later developmental regulation that had not been previously appreciated.

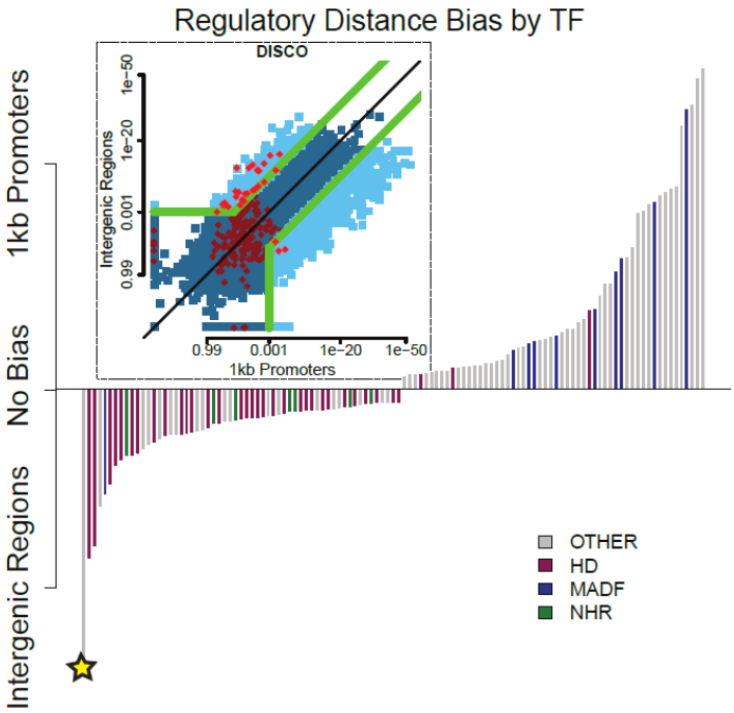


Figure 3.7 Regulatory Distance Bias by TF. Each bar represents a different TF with its color indicting its DBD family and height indicting the statistical strength of the bias between the proximal regulatory region and the more distal, insulator defined regulatory region. The starred transcription factor DISCO is shown in detail in the inset plot with the p-values of the two methods for all TF-domain pairs in blue and for the 195 DISCO-domain pairs in red. Only points outside of the green lines are considered to be significantly biased.

3.7 Applications in other species

Although we have performed our most exhaustive evaluations in the context of *Drosophila* embryonic development, we also applied our pipeline for finding major regulators of co-expressed gene sets to several other species and investigations. These previous applications in other species have employed the same general framework, but lack the incorporation of Brownian Motion multi-species motif scores and chromatin accessibility score filters that are available in the very well-studied *Drosophila*. First, in another study of insect genomes [78],

associations were discovered between motif target gene sets and gene sets defined from Gene Ontology terms. In Figure 4D of that paper, it is reported that almost 40% of the top TF-GO associations recovered with this motif based approach overlap a validation set of ChIP-based associations. This paper also lists a large number of significant TF-GO associations discovered by our approach in both *Drosophila* and *Nasonia* that have strong literature support.

In the songbird genome paper [119], we utilized this method with standard normalization and 5 kbp upstream and 2 kbp downstream regulatory regions on JASPAR [97] and TRANSFAC [98] motifs to analyze differentially expressed genes in the brains of birds exposed to song. In Supplementary Table 6 of that paper, we recovered 12 of 19 motifs from transcription factors that were selected from prior knowledge to have neural activity. We also applied our analysis pipeline on differentially regulated gene sets from four different brain regions and seven separate time points in a recent study of songbird singing [24]. Our analysis revealed that the motifs of early-activated transcription factors that respond quickly to the singing stimuli are enriched in the singing-regulated immediate early genes in multiple brain tissues. In the next chapter, we will present additional regulatory discoveries made with the help of this pipeline in co-expressed sets of genes relating to social behaviors in honeybees, stickleback fish, and mice.

In order to enable researchers to easily employ our method of identification of gene set regulators, we produced an online web tool, MET, for on-demand analysis [83]. Our tool currently functions for a dozen species from flowering plants to bees and from planarian to humans. The tool incorporates several large collections of experimentally characterized TF motifs and enables the multi-species motif scoring and chromatin accessibility filters on well-studied species (i.e. humans and fruit flies). The association results returned by MET are linked to a genome browser of regulatory features that display the regulatory landscape of the putative motif target genes. Finally, the interface is designed to produce regulatory enrichment results in real-time, which enables researchers to explore the several parameters of the pipeline that may affect their results. The address of this webserver is http://veda.cs.uiuc.edu/MET/.

3.8 Discussion

In the evaluations presented in Section 3.3, we demonstrated that computational scoring of motifs is able to predict TF binding profiles. Unlike previously reported methods that trained free parameters from ChIP data, [89, 90, 94], our prediction approach was completely free of hand-tuned parameters. Consistent with our findings in [84] on only 6 TFs, we noted that

evolutionary conservation, measured by a phylogenetically weighted average score of motif presence in orthologous segments, provides substantial improvements in the accuracy of occupancy prediction for dozens of transcription factors. We additionally observe that filtering motif-based computational predictions with cell type-specific accessibility profiles is able to significantly improve the predictions. In this "motif + accessibility" approach, only one experimental assay is needed to study the regulatory landscape of a novel cell or tissue type rather than one assay for every TFs required by ChIP. Additionally, we noted very strong positive and negative correlations between motif presence and accessibility. The informative motifs were often stage-specific, e.g., VFL correlated strongly in the earliest stage analyzed and poorly in the last stage, consistent with its temporal expression profile. Thus, in principle, future methods may be able to utilize expression data on TFs along with their motif profiles to predict approximate accessibility profiles in a stage-specific manner, which then may be utilized to predict stage-specific occupancy profiles for other transcription factors.

Our pipeline relies on finding enrichments with motif target gene sets produced from motif computations scans [35, 120]. This is distinct from *ab initio* motif-finding tools in a few key ways. First, MET implicitly searches the genome for enhancer-like windows that are targeted by a particular transcription factor. Our score for a regulatory feature is not summed over the entire length of a long intergenic regulatory region, but a search for the 500 bp with the strongest regulatory signal in that region. The regulatory region does not have to be the immediate upstream region (e.g., 1 kbp promoters); rather, it may be much longer (e.g., tens of kbp) and is configurable by the user. It would be extremely challenging for standard motif-finding tools like MEME [18] or CONSENSUS [19] to search large regulatory regions for overrepresented motifs whose matches (sites) are localized to one or a few enhancers in the region. Secondly, we provide a more generalized framework than motif-finding methods that associates additional types of regulatory features (chromatin accessibility, chromatin state, TF occupancy) with novel gene sets. Thirdly, a major advantage of our procedure (i.e., enrichment tests with known motifs) over *ab initio* motif-finding algorithms is the reduction of the search space. Motif-finding tools perform a search over large space of possible k-mers (or PWMs) reducing the power of the statistical tests. Moreover, *ab initio* motif discovery is often followed by a post-processing step to relate the identified motif to the most similar known motif. We are

limiting the number of statistical tests by only analyzing experimentally validated TF-DNA binding specificities and thereby potentially increasing the statistical power.

In Section 3.5, we demonstrated our ability to leverage TF motifs to create a large compendium of statistical associations between regulatory TFs and their target tissues and cell type-specific programs. We noted that our motif-based approach has roughly the same accuracy as a ChIP-based approach, again arguing for the proposed alternative paradigm at the heart of this work. With increasing availability of accessibility data, the efficacy of this approach is expected to improve, especially for vertebrate genomes where such data will greatly reduce the search space for *cis*-regulatory signals. The computational pipeline presented here will be particularly useful to biologists who want to understand regulation of genes in non-model organisms or specific cell types that are not investigated by well-funded projects such as ENCODE.

4 SHARED REGULATORY SIGNATURES ACROSS MULTIPLE GENE SETS

This chapter introduces an algorithm for identifying regulatory signal enrichments that are shared across multiple gene sets. The majority of this chapter is taken from a joint work with Seth Ament published in the Proceedings of the National Academy of Sciences [121]. A portion of Section 4.5 is from a collaboration with Prof. Alison Bell's laboratory, published in the Proceedings of the Royal Society B [25].

4.1 Background

The study of the evolution of developmental processes resulted in the observation that underlying the complexity and diversity of animal body plans are a small set of common, highly conserved regulatory components. In fact, many complex phenotypes may be shaped in part by common molecular mechanisms [122]. For example, although a common behavior between two species or metastatic tumors from different primary cancer types may have distinct differential gene expression profiles, both profiles might share an influence by a single sequence-specific transcription factor. Motivated by this observation, in this chapter, we sought to identify a common regulatory signature across multiple, functionally related gene sets derived from different species, tissues, or experimental determinants of a phenotype.

While there are many tools to search for associations with a single gene set, like the work in the previous chapter or popular web tools like DAVID [12] and GSEA[13], there are far fewer that attempt to find shared or "meta-" associations across multiple transcriptomic states. Our approach combines the p-values of multiple association tests into a novel test statistic whose significance can be computed analytically. Unlike most methods for combining p-values from multiple statistical tests, like Fisher's, Stouffer's, and others (reviewed in [31]), our novel statistic is ideal for instances when an unknown subset of the tests are expected to provide evidence against the null hypothesis. This is an extremely useful ability in comparing transcriptomic profiles that might not all share the same molecular underpinnings.

In this chapter, we examine the specific problem of identifying regulatory associations across multiple experimental conditions. We compare our novel framework for common regulator identification in multiple gene sets to basic methods. These alternative methods separately identify core sets of genes ("modules") with shared aspects of their transcriptomic profiles and then subject the modules to follow up regulatory analysis with motif discovery tools like MEME [18]. Biclustering tools like SAMBA [27] and BiMax [26] are designed to identify

these core gene sets that may have similar expression in a subset of conditions. Other methods like COALESCE [29] integrate the two steps described above in an iterative method that identifies modules of genes that simultaneously have a common expression pattern and regulatory motif enrichment. However, these methods are performing *ab initio* discovery of core gene sets and *de novo* discovery of regulatory motifs, two tasks with very large search spaces, and may thus have less statistical power when subjected to rigorous multiple hypothesis correction.

Our method, on the other hand, reverses the two steps of the shared regulator search by first finding associated TF motifs in each condition independently and then combining the multiple p-values into a novel meta-statistic. By only searching with experimentally characterized TF motifs and by defining the allowable relationship between the multiple expression gene sets, we limit the number of statistical tests performed. We also control for the effects of multiple hypothesis testing by comparing our results to empirical extreme value distributions. Another important feature of our novel meta-statistic is that it integrates multiple significance p-values without using any thresholds, unlike methods employed in [8]. Finally, transcriptional regulation in eukaryotes often involves combinations of several transcription factors. Our *cis*-Metalysis tool is able to find user defined, logical combinations of TF motifs that are shared among several gene sets. The *cis*-Metalysis tool provides biologists with a unique ability to discover shared regulatory mechanisms important across multiple transcriptomic states.

4.2 Novel score for combining p-values

Independent gene expression experiments are often performed to better understand the same phenomenon. Each experiment may result in its own set of differentially expressed genes. One approach to identify regulatory associations that are shared across multiple gene sets is to perform independent association tests with each gene set and then combine the significance p-values. The most popular p-value combination approach [30, 31] is the Fisher method, which identifies when at least one of the null hypotheses is rejected and is sensitive to the smallest p-value. Our "meta-p value", on the other hand, identifies when a subset of null hypotheses are rejected. In order to calculate the meta-p value from a list of p-values from n independent tests, we order the list from smallest to largest: $\{p_i\}$. For each $k \in [1..n]$, we compute a meta-statistic:

$$\phi_k = 1 - \prod_{\{i=1\}}^{k} (1 - p_i)$$

which combines the k most significant p-values. Note our meta-statistic is a value between 0 and 1 and is very small only if every p_i for $i \leq k$ is small. We extend the work of [123] to calculate the p-value of the meta-statistic, $P(\phi_k)$, conditional on the fact that the k smallest p-values were chosen from a set of n. This is done analytically with the derived calculation:

$$\Pr(f_k \leq t) = 1 - \int_0^1 \binom{n}{k+1} (k+1)(1-t)^k t^{n-k+1} A(t,k,1-t)\, dt$$

where $A(t,k,x)$ denotes the probability that the product of k independent variables, each of which is uniformly distributed on [t, 1], $\leq \tau$ and is calculated with a function provided in [123].

We find the $\min_k P(\phi_k)$ because we do not assume *a priori* that we know the number of tests that will carry evidence against the null hypothesis and refer to this value the "meta p-value" of a meta-association shared across conditions. An example calculation of the "meta p-value" is worked out in Figure 4.1.

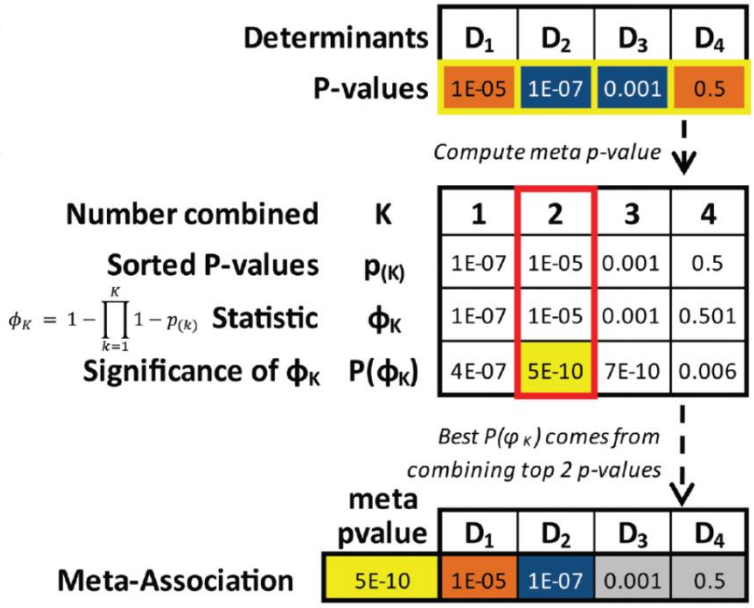


Figure 4.1 Meta P-value Calculation Example. The top table shows the best association p-value from each condition ($D_1 \ldots D_4$). For each K=1…4, the statistic ϕ_K combining the best K p-values is computed and translated to a p-value $P(\phi_K)$ in the center table. The minimum $P(\phi_K)$ over all K is the "meta p-value" (highlighted in yellow within the red border) and considers the number and strength of the combined p-values. The meta-association is represented in the bottom table with selected significant conditions colored and the remaining in gray.

The meta p-value score is ideal for the application in regulatory TF discovery because it considers the number and strength of the combined p-values without assuming the TF was a

regulator in all conditions. One important assumption required is that the combined p-values come from independent tests. In practice, this assumption may be violated; which is why we follow up our meta-analysis with estimations on the false discovery rate.

4.2.1 Comparison of novel statistic to standard method

The popular Fisher's combined probability test calculates the statistic:

$$c_F^2 = -2\sum_{i=1}^{g} \ln p_i .$$

Our meta p-value is designed to be most sensitive to the largest of the *k* best p-values unlike Fisher's method, which is most sensitive to the smallest p-value. In order to compare between Fisher's method and meta p-value in terms of their ability to score meta-associations, we created a synthetic dataset of artificially generated "candidate meta-associations". Each candidate meta-association is an 11-tuple of p-values, whose strength is parameterized by two numbers, *k* and α. The integer *k* is the number of "significant" p-values in the 11-tuple and the real number $\alpha \in [0, 1]$ is a "strength" parameter for choosing those *k* significant p-values. Given (k, α), we randomly generated (11 – *k*) real numbers from Uniform[0, 1], and *k* real numbers (the "significant p-values") from the range [0, α] following an empirical distribution on significant association p-values from real data. We evaluated each candidate meta-association by the meta p-value and the Fisher's combined probability statistic separately and asked which method's significance evaluation correlated better with the parameters *k* and α.

At a fixed value of *k*, we varied α to take eight values (0.00001, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.5, and 1), generated 200 candidate meta-associations for each value of α, computed the test statistic (meta p-value or significance of Fisher's combined probability statistic) on each candidate meta-association, and determined the Spearman's rank correlation coefficient between the statistic and α over all (200 x 8 = 1600) candidate meta-associations. The correlation with our meta p-value statistic was higher than the correlation with the Fisher combined statistic for the tested values of *k* (1, 2, 3, 4). We also found that at the fixed value of α = 0.05, when we varied *k* to take values in [1..11], generated 200 candidate meta-associations for each value of *k*, and determined the Spearman's rank correlation coefficient between the statistic and *k*, our statistic was better correlated than the Fisher statistic (0.91>0.87).

We also attempted to compare our method for selecting the top association p-values in the meta-association to the standard multiple hypothesis correction procedure of using q-values [124]. For five values of α, we created 2200 candidate meta-associations (200 for each value of $k \in [1, \ldots 11]$). Calculation of the "meta p-value" on any candidate meta-association involves choosing an integer *k'* such that $P(\phi_{k'})$ is minimized; this may be adopted as an approach to select a subset of significant p-values from a given set of p-values. We called the integer *k'* the "topk" statistic and determined the correlation between the parameter *k* and the topk statistic. Alternatively, we performed multiple hypothesis correction on the 11 p-values in a candidate meta-association, counted the number of p-values that meet a q-value threshold of 0.05, and called this value the "topq" statistic. We then determined the correlation coefficient between the parameter *k* and the topq statistic. We found that for each value of $\alpha \in \{0.00001, 0.0001, 0.001, 0.01\}$ the "topk" statistic is better correlated with the true number of significant p-values than the more standard approach represented by "topq" (Figure 4.2), suggesting that the meta p-value based approach to select the subset of tests where the null hypothesis was false is better than standard false discovery rate methods.

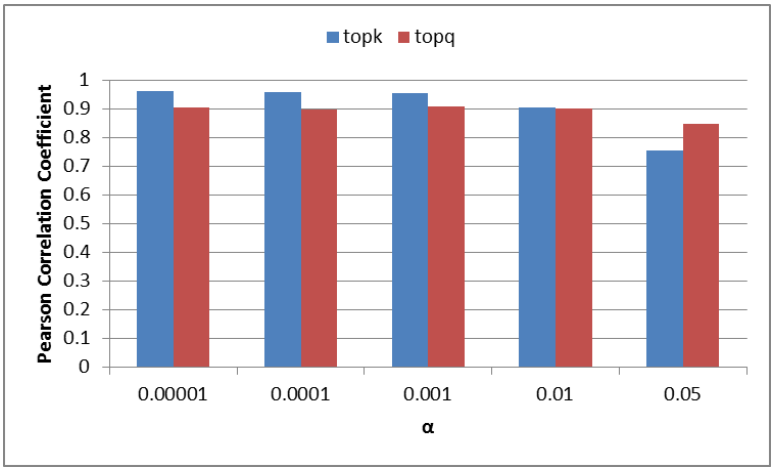


Figure 4.2 Identifying Significant Associations. For each value of α (x-axis), we plot the Pearson correlation coefficients between each statistic (topk or topq) and the true value of parameter k (number of significant p-values) across 2200 synthetic, candidate meta-associations.

4.2.2 Metalysis framework

We created a framework called Metalysis, which systematically searches for significant meta-associations using our novel test statistic. Let *G* denote the universe of all genes, *C* denote the set of experimental conditions for which expression data (on *G*) is available, and *M* denote some collection of annotations for the genes (e.g., several Gene Ontology terms). The Metalysis

program expects two inputs: (i) a "GxM" matrix with a row for each gene in *G*, a column for each annotation in *M*, and binary membership values that indicate that a gene has a given annotation and (ii) a "GxC" matrix with a row for each gene in *G*, a column for each experimental condition in *C*, and 1, -1, 0, and 2 values if the gene in that condition was up regulated, down regulated, not differentially regulated, or not experimentally assayed respectively. We will refer to the genes annotated with annotation property M_k as G_k and the up/down regulated genes of experimental condition C_i as $G_{i,+}$ and $G_{i,-}$ respectively. With these two input matrices, the steps of Metalysis for each possible G_k are: (i) calculate p-values of a Hypergeometric test of association between G_k and $G_{i,+}$ and between G_k and $G_{i,-}$ for each condition C_i, (ii) select the lower p-value p_i for each condition C_i, and (iii) compute the meta p-value from the resulting $\{pi\}$ as the significance of meta-association between M_k and the set of conditions. The current implementation calculates the Hypergeometric tests for quantifying significance because the discretized differential gene sets are most often reported in literature and therefore most widely available. Alternative procedures for calculating significance from continuous expression values (e.g. GSEA [13]) may be substituted into the Metalysis framework.

4.2.3 Multiple hypothesis correction

Since the Metalysis procedure is repeated for each given gene module M_k and since step (ii) amounts to performing two tests for each C_i, a multiple hypothesis correction is required. In order to quantify the quality of meta-associations, we examine the outcomes of random permutations of the real data. The gene labels of the GxM matrix of annotation membership are randomly shuffled. The entire analysis is repeated on the permuted data and the most significant meta p-value reported by Metalysis is recorded. We repeat this exercise many times and construct an empirical extreme value distribution (EVD) of meta p-values. We then approximate the empirical EVD by fitting a Gamma distribution to it, as has been reported previously in the context of *ab initio* motif discovery tools [125]. We examine this smooth distribution to calculate an "EVD p-value" corresponding to each meta p-value in the original dataset. We set thresholds on the EVD p-value to control for multiple hypothesis testing and return only the most reliable significant meta-associations. This is a very conservative form of multiple hypothesis correction because it places a threshold on the chance of finding *any* meta-association of a given significance.

4.3 *cis*-Metalysis framework for identifying shared regulators

We have developed a specific version of the Metalysis framework to rigorously test for shared regulatory signatures across gene sets from multiple experiments. It relies on the scans of gene promoters for transcription factor DNA-binding motifs to predict which genes may be regulated by that TF, and then performs meta-analysis using this information. In this *cis*-Metalysis framework, we define motif modules using the same techniques described in Chapter 3. However, because TF regulation in eukaryotes is often combinatorial in nature, we extend our method and additionally create new motif modules defined from logical combinations of the presence and absence of TF motifs. Moreover, it can be configured so that different logical combinations of the same motifs may be associated with DEGs in different experiments, thereby offering a flexible model of regulatory mechanisms shared by multiple transcriptomic states. Figure 4.3 provides an overview of the *cis*-Metalysis pipeline.

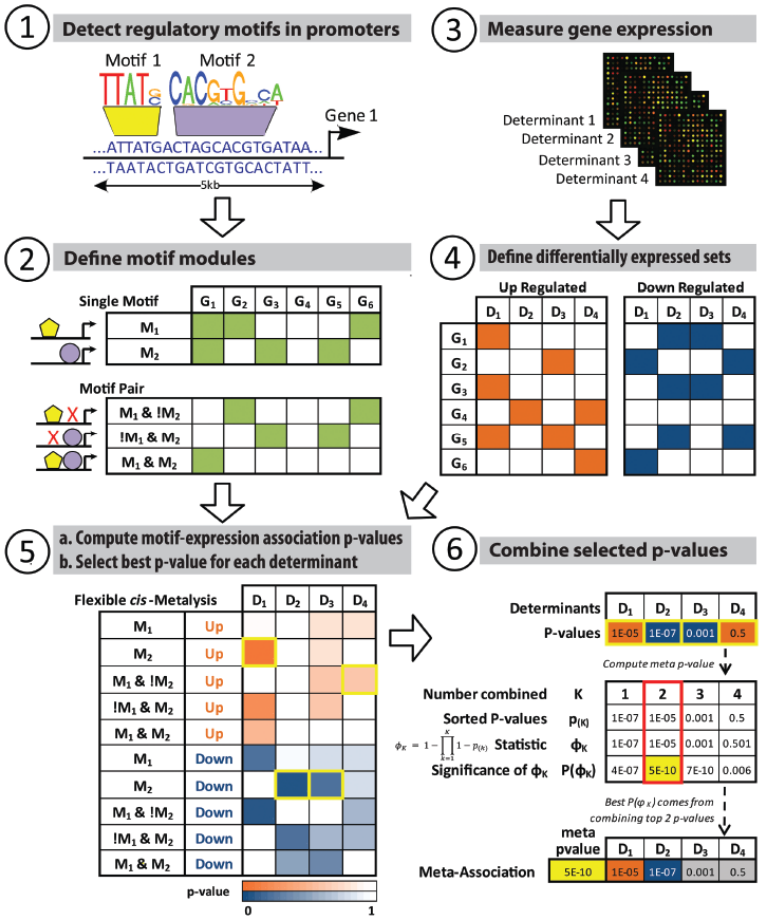


Figure 4.3 Overview of cis-Metalysis. Steps 1 and 2: Motif modules are defined based on the presence (green cells) of single motifs e.g., M_1, or their Boolean combinations, e.g., M_1 & M_2 in the 5 kbp promoter sequences of each gene. Step 3 and 4: Sets of up- (orange) and down- (blue) regulated genes and identified from experimentally profiling gene expression for each determinant. Step 5: Statistical enrichments are conducted between motif modules

and expression sets producing all of the motif-expression association p-values (shaded cells) that are combined by *cis*-Metalysis. In its most flexible mode, *cis*-Metalysis attempts to combine the best p-value (bordered in yellow) per determinant. Step 6: Calculation of meta p-value test statistic.

4.3.1 Modes of *cis*-Metalysis

An important component of *cis*-Metalysis is the different ways of defining the *cis*-regulatory logic shared by multiple conditions. Particular hypotheses may be examined depending on whether different experimental conditions must be associated with rigidly or flexibly defined regulatory modules. There are five distinct modes of *cis*-Metalysis, which the user must specify at runtime. The "Single motifs" mode of *cis*-Metalysis is an instance of the general Metalysis framework. Each motif module is treated as a separate motif module and examined for its own meta-associations. No combinatorial motif modules are created in "Single motifs" mode. The "Identical logic" mode enforces the most rigid definition of regulatory modules involving multiple motifs. For any motif pair (m_1, m_2), the combinations $m_1 \wedge m_2$, $m_1 \wedge \neg m_2$, and $m_2 \wedge \neg m_1$ are analyzed separately by constructing the respective combined motif modules from the motif modules of m_1 and m_2. Each meta p-value significance is calculated for each meta-association between one of the three derived modules and only the up regulated ($G_{i,+}$) or down-regulated gene set ($G_{i,-}$) in every experimental condition C_i. The best meta p-value among the motif combinations is reported for the motif pair. In the "identical logic" mode, meta-associations are reported when the same motif signature has the same differential regulatory effect in multiple conditions. The third mode, "Role consistent logic", requires that meta-associations involving multi-motif regulatory modules use the component TFs in the same regulatory role (effective activator or repressor of gene expression). For any motif pair (m_1, m_2), associations with different conditions may involve any of the following motif modules: m_1, m_2, $m_1 \wedge m_2$, $m_1 \wedge \neg m_2$, and $m_2 \wedge \neg m_1$; however all associations must be mutually "role consistent" in the sense that if m_i is associated with up-regulated genes in one condition, then m_i may not be associated with down-regulated genes in another condition, nor may $\neg m_i$ be associated with up-regulated genes in any condition. Figure 4.4 shows an example meta-association derived from the combinatorial rules of each of these modes.

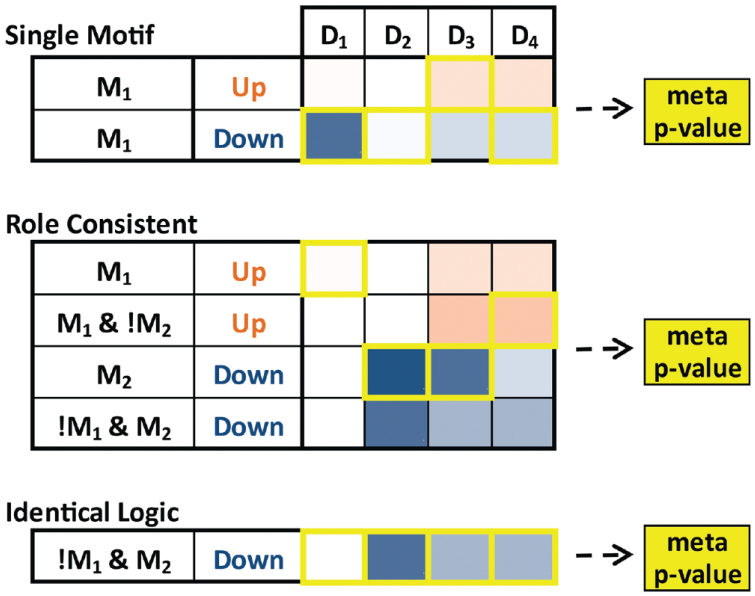


Figure 4.4 Modes of cis-Metalysis. Multiple modes of *cis*-Metalysis enable discovery of simple or combinatorial, identical or plastic forms of regulatory logic shared across transcriptomic states (conditions) by selecting specific subsets of association p-values to combine. For each mode, an example subset of associations (shaded cells) as well as the selected best p-values (bordered in yellow) is depicted. In "single motif" *cis*-Metalysis, p-values are selected from the association tests with the gene modules defined by a single motif (M_1). In the "identical logic" configuration, a rigidly defined combination of two motifs ($!M_1$ & M_2 here) defines the gene module that is tested for association with a fixed direction of regulation ("down" in example shown). In "role consistent", a meta-association of a motif pair (M_1,M_2) is allowed to take different Boolean combinations of the two motifs, and different directions of regulation, in different determinants. However, in "role-consistent" mode, each motif-expression association must use a particular motif in the same "role" (activator or repressor); here, M_1 is an effective activator and M_2 a repressor.

The "Flexible logic" mode allows for any combination of motif module combinations and differential expression direction combinations across experimental conditions. Specifically, for any motif pair (m_1, m_2), associations with different conditions may involve any of the following motif modules: m_1, m_2, $m_1 \wedge m_2$, $m_1 \wedge \neg m_2$, and $m_2 \wedge \neg m_1$. No further constraints are imposed here in defining a valid meta-association. Panel 5 of Figure 4.3 shows a meta-association defined with "flexible logic". The final mode, "Pattern logic" is able to test very specific hypothesis that involve the relationship between experimental conditions. For example, if two different tissues have opposite responses in an experiment, there might be an important relationship between the up-regulated genes in one tissue in the down-regulated genes of the second. "Pattern logic" will require that the expression directions of the motif-expression associations that comprise the meta-association follow the restrictions set by the pattern. The source code for the *cis*-Metalysis program is available for free download at http://veda.cs.uiuc.edu/cisMetalysis/.

4.4 Application to nursing and foraging behavior in honeybees

We have successfully applied our new statistic and framework to a number of biological systems. Our initial and most detailed application of the *cis*-Metalysis framework was to study the determinants of honeybee maturation. Honeybees perform tasks inside the hive for the first 2-3 weeks of their adult life and then switch into roles of foraging for food outside. Many known determinants delay or accelerate this behavioral maturation including pheromones, nutrition, and genetic factors. Our task was to identify a shared regulatory signature across the differentially expressed genes of subsets of eleven maturation determinants. The 11 maturation determinants and the relationships between them are summarized in Figure 4.5.

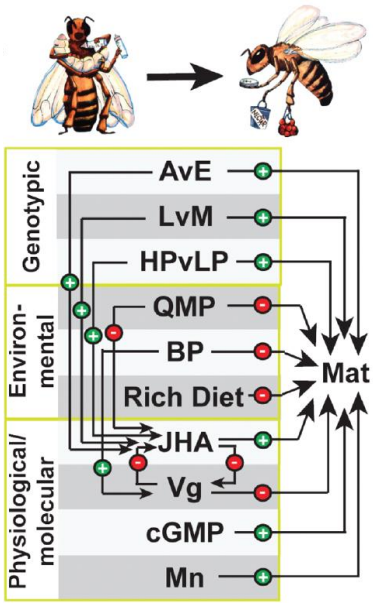


Figure 4.5 Determinants of behavioral maturation. The 11 maturation determinants in this study are listed within the yellow boxes representing different classes of maturation. "Mat" represents maturation (nurses vs. foragers). For genetic comparisons, Africanized vs. European sub-species (AvE), Northern (*A. mellifera mellifera*) vs. Southern European (*A. mellifera ligustica*) sub-species (LvM), and high vs. low pollen-hoarding genetic strains (PH), the first genotype shows faster maturation. Environmental factors like Queen Mandibular Pheromone (QMP), brood pheromone (BP), and rich vs. poor diet ("Diet") also affect maturation. Finally, chemical determinants of maturation include *vitellogenin* RNAi (Vg), juvenile hormone analog treatment (JHA), manganese treatment (Mg), and cyclic-guanosine monophosphate treatment (cGMP). Known stimulating or inhibiting relationships between maturation determinants are represented by "+"or "-" arrows respectively.

In this study, microarray experiments on nearly 400 bees provided transcriptomic profiles. For each maturation determinant, 100s-1000s of differentially expressed genes were identified. The genes that were more highly expressed in the faster or slower maturing bees were referred to as the "fast" or "slow" maturation genes and took the place of "up" and "down" regulation labels in the Metalysis framework. These "fast" and "slow" gene sets defined the required GxC matrix with one column for each maturation determinant. Similarity between the

determinants was quantified by comparing their gene expression profiles; however, this analysis did not reveal any insights into shared gene modules that might underlie the common phenotypic effect of distinct maturation determinants. The application of Metalysis and *cis*-Metalysis enabled us to identify these types of relationships and quantify their significance.

4.4.1 Results of Metalysis and *cis*-Metalysis

First, we employed Metalysis to identify Gene Ontology [10] defined gene modules that relate to behavior maturation. We focused on 613 biological process GO terms with between 10 and 1000 annotated genes in *D. melanogaster*. We mapped the annotations of these terms to their *A. mellifera* orthologs to construct the GxM matrix with 613 annotation columns required by Metalysis. Metalysis was applied to find meta-associations between each GO gene set and the differentially regulated gene sets from subsets of the 11 maturation determinants. We found biological processes enriched in as many as 8 of 11 maturation determinants (Table 4.1). These meta-associations included processes occurring in the brain related to macronutrient and energy metabolism (translation, mitochondrial electron transport, glycolysis), neuronal plasticity (synaptic transmission, nervous system development), and stress responses (protein folding, response to heat). The EVD p-value threshold of 0.05 was applied to guarantee the significance of the reported results.

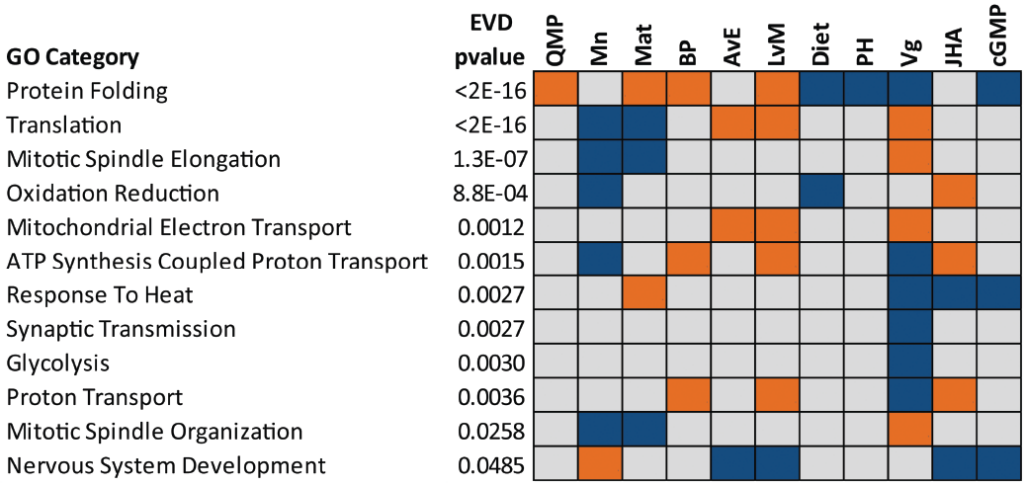


Table 4.1 Metalysis Results with Gene Ontology. Significant meta-associations between Gene Ontology biological processes and maturation determinants are reported with their EVD p-value. Colored cells indicate the individual maturation determinants was included in the meta-association with orange (resp., blue) denoting an association with "fast" (resp., "slow") maturation genes.

Next, we wanted to examine the hypothesis that multiple maturation determinants operate through the actions of a common set of TFs. To do this, we employed the various modes of *cis*-

Metalysis to explore both simple and complex models of regulation. To construct the GxM matrix, we needed to find the motif modules for each of 602 TF motifs that were downloaded from multiple sources [97, 98, 104]. These motif modules, originally created in [126], are defined with a method similar to our approach in Chapter 3. We searched up to 5 kbp upstream of a gene for a motif's presence using the SWAN program [78], which captures the presence of one or more, strong or weak matches to the motif in the genomic segment, and accounts for the local G/C composition as well as the global frequency of motif occurrence. The motif modules defined the GxM Boolean matrix with 602 columns and values representing if the TF binding motif is present in the gene's upstream region. Using the "single motif" mode of *cis*-Metalysis, we found meta-associations for 22 motifs that spanned four to six of the determinants. We then tested the "Role-Consistent" mode of *cis*-Metalysis to find motif pairs with significant meta-associations. This analysis returned meta-associations (Table 4.2) involving up to ten maturation determinants and involving motifs for well-known TF regulators of neuronal plasticity (CREB) and stress response (XBP1). Finally, when allowing for the "flexible logic" mode, we identified 16 meta-associations involving 20 TFs with very significant EVD p-values (< 2E-16) and spanning all 11 of the maturation determinants. These results suggest that many different maturation determinants use the same TFs to exert common effects on behavior, but that different determinants employ some of them in distinct ways.

Motif 1	Motif 2	EVD pvalue	QMP	Mn	Mat	BP	AvE	LvM	Diet	PH	Vg	JHA	cGMP
CREB1	Runt	<2E-16											
ETS	XBP1	1.5E-03											
ETS	Dl	3.8E-03											

Table 4.2 Top Role Consistent Meta-Associations. Top three most significant meta-associations between maturation determinants and pairs of motifs that interact with role-consistent logic. Orange (resp., blue) denotes that the motif combination was associated with "fast" (resp., "slow") maturation genes in that experiment. Motifs whose presence is associated with "fast" and with "slow" maturation genes are in green and red font, respectively.

Several of the TFs (including CREB, BR, DL, XBP1) identified by *cis*-Metalysis have also been implicated as high-level regulators of maturationally related gene expression in a bee brain transcriptional regulatory network reconstructed from gene expression datasets [127]. Additionally, we identified meta-associations involving *ultraspiracle*, a JH-related TF that has been demonstrated with ChIP-ChIP to bind near maturation-related genes, and whose RNAi knockdown has been shown to delay the onset of foraging behavior [6].

4.4.2 Comparison to other methods

Our final analysis on the honeybee maturation data was a comparison of *cis*-Metalysis to alternative algorithms that depend on biclustering, a procedure for finding the subsets of genes and experiments such that the chosen genes are coordinately expressed in the selected experiments. Once such a co-expressed gene set has been discovered, it is common to test for *cis*-elements overrepresented in their promoters, thereby inferring meta-associations between *cis*-elements and the selected experiments. To test this strategy, we applied the BiMax tool [26] to find biclusters across the 11 maturation determinants. 492 biclusters were discovered, each including at least five genes and spanning at least three determinants. The genes in each of these biclusters were then tested for enrichment of motifs in our collection of 602 motifs. We then repeated the entire analysis 50 times on randomized datasets, exactly as described for Metalysis, to obtain empirical EVD p-values that correct for multiple hypothesis testing. We repeated the above analysis with another biclustering tool called SAMBA [27]. Here we input the log2 fold-change values of genes, rather than their discretization into one of three categories (up, down, or neither) and the default SAMBA parameters. Twelve biclusters were discovered that contained between 19 and 90 genes and covered 4 to 5 maturation determinants. We subjected these biclusters to motif enrichment tests and performed the EVD analysis with 50 randomized datasets. Figure 4.6 shows the empirical EVD from each of the 50 negative controls for meta-associations discovered with single motif *cis*-Metalysis, the method using BiMax biclusters, and the method using SAMBA biclusters. This figure shows that *cis*-Metalysis is best suited for finding meta-associations that are statistically significant (i.e. to the right of the empirical EVD). The relative lack of statistically significant meta-associations in the BiMax and SAMBA analyses were also observed for combinations of motifs (not shown).

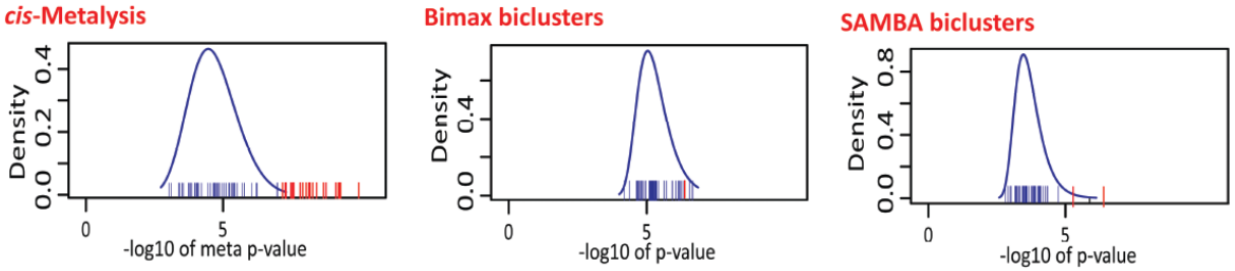


Figure 4.6 Comparison to Biclustering Methods. The empirical EVD distribution from the runs of each method on 50 randomized datasets is represented with the blue tick marks and its fitted Gamma distribution is represented by the blue curve. Each red tick mark represents a meta-association found on the real, non-permuted data with an EVD p-value less than 0.01.

4.5 Application to other systems

The flexible and powerful statistical framework of *cis*-Metalysis leads to its broad applicability. As an illustrative application of *cis-Metalysis,* we attempted to identify potential regulatory signals that underlie the development of cancer. We extracted thirteen sets of differentially expressed genes involved in breast cancer from the curated gene set collection of MSigDB [13]. *cis*-Metalysis was run on the gene sets defined by these thirteen cancer studies with 432 motif target modules defined from scanned human gene promoters as described in [128]. We found thirteen TF motifs with significant meta-associations (EVD p-value < 0.05) spanning at least two studies. We also conducted a literature survey and discovered that most of our identified TFs have been previously linked to breast cancer (Table 4.3). We repeated the procedure separately for 6 kidney and 10 liver cancer gene sets and discovered 16 and 15 significant meta-associations respectively.

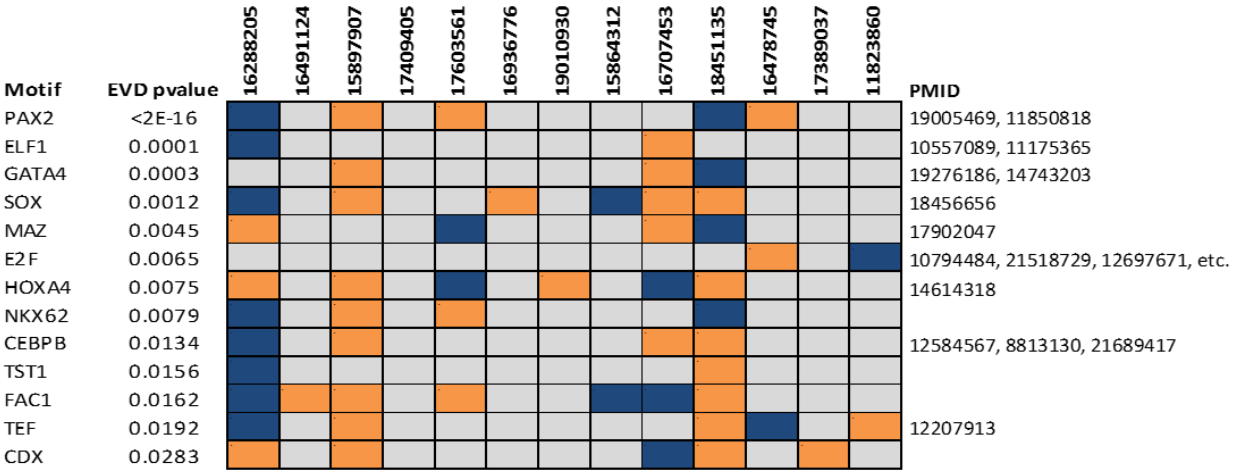


Table 4.3 cis-Metalysis Results on Breast Cancer Gene Sets. Significant meta-associations (EVD p-value < 0.05) spanning at least two studies were identified for thirteen motifs. Each column in the above table represents one of thirteen breast cancer gene sets labeled with the PubMed ID of its source publication. The orange (resp., blue) cells indicate that the motif (row) is enriched in up- (resp., down-) regulated genes from the breast cancer study (column). The PubMed ID of each publication discovered in our literature survey that suggests a relationship between the motif's corresponding transcription factor and breast cancer is reported in the rightmost column.

In another study [25], we applied our *cis*-Metalysis tool to the differentially expressed gene sets from multiple brain tissues produced in response to a behavioral stimuli. The male stickleback fish has a well-characterized aggressive behavior when other male fish intrude their nesting territories. In this study, microarrays characterized the transcriptomic profiles of four distinct brain tissues (diencephalon, telencephalon, cerebellum, and brain stem) of aggressive fish 30 minutes after exposure to an intruder. We constructed a four-column GxC matrix for *cis*-Metalysis using the differentially regulated genes in the four brain regions. Our GxM matrix was

constructed with 661 TFs motif modules using the techniques described in Chapter 3. We found significant meta-associations for 13 TF motifs using "single motif" *cis*-Metalysis. Many of these were consistently associated with one direction of differential regulation across all or most brain regions (NRF2, RREB1, PPARG, POU3F2, etc.). In the examination of the gene expression evidence, one of the most striking observations was that many genes that were upregulated in the diencephalon were down regulated in all other brain regions. We applied the "pattern logic" mode of *cis*-Metalysis and identified several significant motif pairs (NRF2/ER, BACH2/LMO2COM, and NRF2/SRF). These motif pairs are part of meta-associations in which the TF motifs are associated with the up-regulated genes in diencephalon and the down-regulated genes in the other brain regions.

The *cis*-Metalysis framework is especially useful in studying co-expressed gene sets of multiple species. It was applied to identify common elements of the transcriptional regulatory network in a study that examined the brain's response to social challenges in species as diverged as the honeybee, stickleback, and mouse [7]. It found the nuclear receptor "toolkit" TF NR2E1 enriched in all species and several other TFs enriched in two of the three species, including the neuroendocrine signaling NRFA, mentioned above.

4.6 Discussion

We solved a fundamental statistical problem in meta-analysis by developing informatics tools that analyze sequence and expression data to reveal a flexible *cis*-regulatory code underlying a complex phenotype. We employ a meta-analytic strategy where multiple assays of the same experimental condition are analyzed first, gathering robust information on differentially regulated genes of that condition, which is then integrated across multiple conditions. This is in contrast to strategies devoted to finding core set of genes with coordinated expression in multiple experiments and then examining for their regulatory signature. Our approach relies on a more fundamental notion that it is not a list of common genes that are shared across multiple transcriptomic responses, but a common biological process or a common regulatory logic.

Our findings may reflect a significant theme in the regulation of complex behavior phenotypes. The observation that multiple determinants of the same phenotype utilize common regulatory components is not surprising and is reminiscent of the diverse mechanisms for development which rely upon a common toolkit of regulatory genes [129]. Less appreciated is the possibility that these common components may be wired differently in the different

regulatory networks, reflecting the different adaptive forces that shaped the evolution of each of those networks. This possibility poses significant challenges to characterizing an underlying regulatory code for the phenotype, which *cis*-Metalysis attempts to overcome.

Our approach has some parallels with [8], where reported gene modules exhibit associations with several transcriptomic states (cancer types) that have a common annotation (e.g., metastatic cancer). However, their approach requires *a priori* annotation of the transcriptomic states into two categories, and thus cannot reveal meta-associations in a framework such as the one described here. Moreover, its meta-analytic statistics count the number of significant associations at an arbitrary threshold, as opposed to Metalysis, which integrates the strengths and number of associations without imposing thresholds. In addition, their approach does not report gene modules representing combinatorial *cis*-regulatory codes.

A potential limitation of the work presented here is that *cis*-regulatory analysis was based on analysis of promoter regions, whereas regulatory information in metazoan (especially vertebrates like humans or stickleback fish) is frequently located more distally. We note however that our definition of a "promoter" is the region 5 kbp upstream of a gene, which is expected to include a substantial fraction of regulatory elements. As discussed in Chapter 3, in a genome where additional clues about regulatory locations, such as DNA accessibility and co-factor binding, are available, one may be able to scan more comprehensively for motif matches. We also note that *cis*-Metalysis may be run with any definition of the sequence space and any collection of characterized TF motifs, not just the particular choices made here. We have implemented *cis*-Metalysis to be able to efficiently explore a large motif collection with many options for controlling the complexity of the uncovered meta-associations. Finally, we anticipate that future improvement to our approach presented in this chapter will incorporate information about TF expression levels when predicting regulatory roles for motifs. We will show the value of this additional information source in the next chapter in modeling the transcriptional effect of regulatory sequences in *Drosophila* development.

5 MODELING AND ANNOTATING CELL TYPE SPECIFIC ENHANCERS

This chapter describes enhancer activity models developed for the annotation of enhancers in embryonic fruit flies. The study on anterior-posterior segmentation is part of a joint work with Majid Kazemian that was published in PLoS Biology [84].

5.1 Background

A central challenge in understanding metazoan genome sequences is to identify and annotate the enhancers that regulate the complex spatial and temporal patterns of gene transcription. Recent progress in identifying enhancers has been made with the advent of high throughput experimental assays that measure the state of cell type specific chromatin [23, 130]. There are also experimental methods that enable the quantification of gene expression driven by enhancer elements mostly by creating and incorporating reporter constructs into the genome that contain the sequence near either a fluorescent gene or transcribed barcode (reviewed in [32]). However, the task of assigning the cell type specific regulatory output of each enhancer to specific facet(s) of the gene expression activity of its target gene(s) is largely unsolved. This is complicated by the fact that enhancers may regulate multiple, non-neighboring genes. Spatial organization maps of chromatin [131] may help identify the enhancer's target gene, but the sequencing requirements of the technology make it prohibitive for most applications.

It has been suggested that while dynamic chromatin states paint broad brushstrokes of the regulatory landscape, transcription factors help set up more nuanced, cell type-specific expression programs [72, 132]. Computational annotations of enhancers typically rely on discovering a heterogeneous cluster of TF binding sites (matches to their DNA-specific binding motifs); however, these approaches often result in many falsely annotated enhancers. Thus, an alternative strategy for assigning enhancer driven expression activity may rely on binding potential and expression of the enhancers' regulatory TFs. Successful models of enhancer activity have been constructed using TF binding data from ChIP experiments on relevant transcription factors [38-40, 60, 133]. Others have created enhancer activity models from motif based computational approaches to infer binding and then gene expression [44, 45, 48, 134]. These enhancer models have been based on machine learning approaches like SVMs or thermodynamic based systems [44, 45]. Our approach for enhancer activity modeling incorporates TF expression and predicted TF binding strength in a simple regression framework that provides a simple understanding of the role of the transcription factor and enables the

application of the model to sequences throughout the genome. Since a single enhancer may only drive a discrete aspect of the full expression pattern of its gene, we developed a novel score of enhancer activity prediction that rewards sequences that drive a subset of its neighboring gene's expression pattern. We apply our enhancer activity model to produce a detailed transcriptional regulatory network of the anterior/posterior segmentation in *Drosophila* embryos. We demonstrate how our enhancer annotation approach allows additional insights into how multiple enhancers contribute to gene expression patterns and how individual TFs directly or indirectly regulate the expression of multiple target genes.

Most of the previously mentioned enhancer activity models have been limited to very few, well-characterized cell types and regulatory networks. This is because these approaches require significant prior knowledge in the form of the relevant TFs, genetic knockdowns, validated enhancers, etc. We sought to understand how well our approach would apply to the many tissue types where the available data is limited to the genomic sequence, the tissue specific gene expression, and experimentally assayed chromatin accessibility. Like [23], we employ accessibility to initially identify putative enhancers. Then for each cell type, we build an enhancer activity model that incorporates computational prediction of relevant transcriptional regulators (Chapter 3), their TF binding score in the enhancers and their gene expression, along with the chromatin accessibility of the enhancers. Without using any prior knowledge to train models, we were able to accurately recover enhancers for over 50% of our evaluation cell types and outperform ChIP-based models. Finally, we annotated ambiguous enhancers for the likely expression pattern they produce, finding a large number of distal cell type specific regulatory sequences [135].

5.2 Enhancer modeling in segmentation system of *Drosophila* embryos

Our first goal was to successfully model the 46 well-studied enhancers involved in the anterior-posterior (A/P) segmentation of the blastoderm stage *Drosophila* embryo [50] that have been characterized with reporter gene assays. The pattern of gene expression driven by each of these enhancers along the A/P axis was represented by a binary expression value for 100 bins along the axis (bin 1 is most anterior and bin 100 is most posterior. The enhancers and genes in the A/P system have been well studied (see Chapter 2) and the 10 transcription factors (BCD, CAD, HB, KNI, KR, GT, HKB, TLL, FKH, and CIC) are believed to be important for the formation of the proper segmentation in the embryo. For these 10 transcription factors, we found

the DNA binding specificities ("motifs") as characterized by the bacterial-one hybrid assay [75]. For each of the ten motifs, we produced the multi-species Brownian Motion scores from 11 *Drosophila* genomes as is described in Chapter 3. We also extracted *in situ* images of the 10 transcription factors [50, 136] to converted their blastoderm stage expression patterns ("concentrations") into 100 bin representations with values between 0 (for no expression) and 1 (maximum expression).

5.2.1 Model construction and evaluation

We developed a simple regression based model that captures the expression driven by an enhancer in a bin as a function of (i) each TF's multi-species motif score in the enhancer's sequence and (ii) each TF's concentration value at that position bin. Specifically, we employed a logistic regression model, which has the desirable property of constraining the minimum and maximum activity for all enhancers to 0 and 1 respectively. The parameters of the model include a coefficient representing each TF's regulatory effect and a baseline expression value for each enhancer (which is constant across all bins). This separate parameter for each enhancer is motivated by (i) the fact that the discrete (0/1) expression values that form the desired output do not reflect the variation in basal gene expression levels and (ii) an opportunity to compensate for, at least partially, the lack of complete knowledge of relevant TFs, especially of ubiquitous activators and/or repressors. These parameters were trained on the known expression profiles from the 46 enhancers simultaneously.

The basic model for predicting enhancer expression patterns is as follows:

$$E_{l,b} = sig\left(w_0^l + \sum_{i \in tf} w_i \gamma_{ib} C_i^l \right)$$

where

- E_{lb} is the expression value (between 0 and 1) of the enhancer *l* in bin *b*
- γ_{ib} is the concentration of TF *i* in bin *b*,
- C_i^l the is the multi-species motif score of TF *i* in the enhancer *l*,
- w_i is the regression coefficient for TF *i*
- w_0^l is the "basal" expression level of enhancer *l*
- *sig*(x) is a "sigmoid" function 1/(1+exp(-x)).

We additionally included a higher order term, called "BCD2", in our model. BCD2 is the square of the covariate "BCD" for the factor BCD. Utilizing the glm (generalized linear model) function in R's "stats" package [137], we trained the parameters of the model using iteratively reweighted least squares (IWLS) to minimize the error between predicted and true expression values. The overall quality of fit of the model to the data was measured by standard statistics such as the root mean squared error (RMSE), average Correlation Coefficient (CC), and the Akaike Information Content (AIC).

The fitted model provides "systems level" insights into the A/P network. Overall, it captured 20, 15, and 11 of the 46 enhancers well, fairly, or poorly respectively, with an average correlation coefficient across all enhancers of 0.48. We observed that coefficients for BCD, CAD, and FKH were fit to positive values, while KNI, KR, GT, HB, TLL, HKB, and CIC were fit to negative values, which is broadly consistent with the activator/repressor roles known for these factors. (Although dual roles for some of these factors have been noted in the literature [138], our model learned a single dominant role consistent with the dataset.) Of several "second order" terms we explored, only the one for BCD significantly improved the model (Table 5.1). This BCD2 term produced a broad anterior dip in the BCD concentration gradient and may reflect that our model may not completely account for some aspect of down regulation of BCD target genes by the terminal patterning system, either by converting BCD into a repressor [139] or through regulation of other repressors [140, 141]. We also found that the enhancer activity model that incorporates multi-species motif scores outperforms the one with single species scores. This fact is broadly consistent with previous studies demonstrating that A/P enhancers with conserved activity patterns and similar binding site composition are identifiable in related species [142, 143]. For the eight transcription factors for which ChIP data is available [71, 107], we replaced the motif score profiles with ChIP scores, and retrained the regression model using these data. By statistical measures, the overall quality of fit of the ChIP-based model was inferior to that with multi-species motif profiles, suggesting in the context of this experimental system, comparative genomics may have equal or even greater utility than ChIP-based measurements of TF occupancy (Table 5.1).

Model Implementation	RMSE	Avg. CC	AIC
Single Species (without BCD[2])	0.3135	0.43	3028
Single Species (with BCD[2])	0.3088	0.46	2962
Multi-Species (simple Averaging)	0.309	0.47	2966
Multi-Species (BM Averaging)	0.3046	0.48	2894
ChIP-chip	0.3162	0.36	3109

Table 5.1 Evaluation of A/P Enhancer Model. Evaluation of different variants (column 1) of the logistic regression model, using three different goodness of fit measures: RMSE, Average CC, and AIC.

5.2.2 Annotating putative enhancers

We next employed our fitted regression model to identify novel enhancers by scanning the flanking genomic sequences of a gene for segments whose predicted activity pattern agrees with the gene's endogenous pattern. For this purpose, we developed a new measure of similarity between our 100 bin expression profiles called the "Pattern Generating Potential" (PGP). The scoring measure was designed to: 1) be sensitive to both the shape and magnitude of the predicted expression profile, 2) avoid biases towards or against overly broad or overly narrow domains of expression, and 3) automatically select aspects of a gene's expression pattern to be captured by the enhancer. Figure 5.1 visualizes these desirable properties.

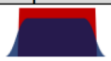
Characteristic	Expression	PGP	CC	1-RMSE
Sensitive to scaling		0.81	0.96	0.81
		0.58	0.96	0.43
Sensitive to shift in basal expression		0.69	0.96	0.62
		0.39	0.96	0.71
Normalization for length of expression domain		0.69	0.96	0.62
		0.69	0.94	0.74
Sensitive to partial pattern		0.56	0.40	0.53
		0.51	0.63	0.62

Figure 5.1 Properties of PGP. Design features of the PGP score that distinguish it from the correlation coefficient (CC) or the root mean square error (RMSE). For each desired feature ("Characteristic"), two scenarios of comparison between known (red) and predicted (dark blue) expression profiles ("Expression"), along with PGP, CC, and 1-RMSE values are shown. A perfect match would correspond to a value of 1 for each score. Cases where the value of a score in the two scenarios captures the desired feature are shaded in green.

Given a predicted expression profile (real numbers between 0 and 1 for each bin along A/P axis) and an endogenous expression profile (0 or 1 values for each bin), we defined the PGP score as follows:

$$PGP = 0.5 \times \left(1 + \frac{\sum_{b \in bins} E_{g,b} \times \hat{E}_{g,b}}{\sum_{b \in bins} E_{g,b}} - 3 \times \frac{\sum_{b \in bins} (1 - E_{g,b}) \times \hat{E}_{g,b}}{\sum_{b \in bins} (1 - E_{g,b})} \right)$$

where $E_{g,b}$ is the expression value (0 or 1) of the gene g in bin b and $\hat{E}_{g,b}$ is the predicted expression value (between 0 and 1). This score ranges from -1 to +1. It rewards correctly predicted domains of expression and penalizes false prediction of expression. If the endogenous profile has multiple domains of expression, a subset of those domains are selected based on the predicted profile and then compared to the predicted profile using PGP.

With the PGP score able to correctly identify expression predictions that capture discrete aspects of a binary A/P expression pattern, we designed a method for finding putative enhancers of A/P genes. We began by obtaining 100 bin A/P expression profiles for the 22 genes regulated by our set of 46 enhancers from data obtained BDGP [113] and FlyExpress [144]. We scanned the regulatory region of each gene (starting from 10 kbp upstream of the gene until 10 kbp downstream) with a sliding window of size 1 kbp. We extracted the multi-species motif score and applied our fitted model to predict the A/P expression profile of that window. Finally, we calculated the PGP score using the predicted enhancer activity and the known gene expression. We calculated an empirical p-value for each PGP score estimated based on how frequently we observed a window with equally high PGP score when scanning genome-wide. Of the 62 significant modules predicted, 34 overlapped (>50%) with known enhancers, indicating our approach has 55% specificity at 74% sensitivity. Seventeen of the remaining 28 predicted modules overlapped the bound regions of at least one transcription factor (ChIP data at 1% FDR from [71, 107]), suggesting that the majority of predicted enhancers are functional and/or biochemical targets of A/P factors. The 12 known modules not recovered included 10 that were not predicted well by the original regression model. The genomic location and predicted expression activity for each of these enhancers are available at http://veda.cs.uiuc.edu/lmcrm.

Unlike the other enhancer prediction approaches, the PGP method predicts which aspect of the gene's pattern is regulated by an individual enhancer, allowing the range of regulatory architectures for the A/P-22 genes to be examined: solitary enhancers, multiple enhancers contributing to distinct aspects of the pattern, or multiple "sibling" enhancers with a similar predicted activity (Figure 5.2). In all but one gene (*btd*), two or more regulatory modules were predicted in a single gene's control region. These included cases where distinct aspects of a

gene's activity are captured by distinct predicted enhancers (e.g., five enhancers near the gene *eve*, including four known enhancers), a well-established phenomenon reported for primary pair-rule genes. We also found several cases of "sibling" enhancers, where multiple modules near a maternal/gap gene were predicted to drive highly similar expression patterns. Given the previous identification of "shadow" enhancers in the dorsal-ventral patterning network [145], the utilization of functionally similar enhancers may be a more common theme of *cis*-regulatory organization than currently recognized.

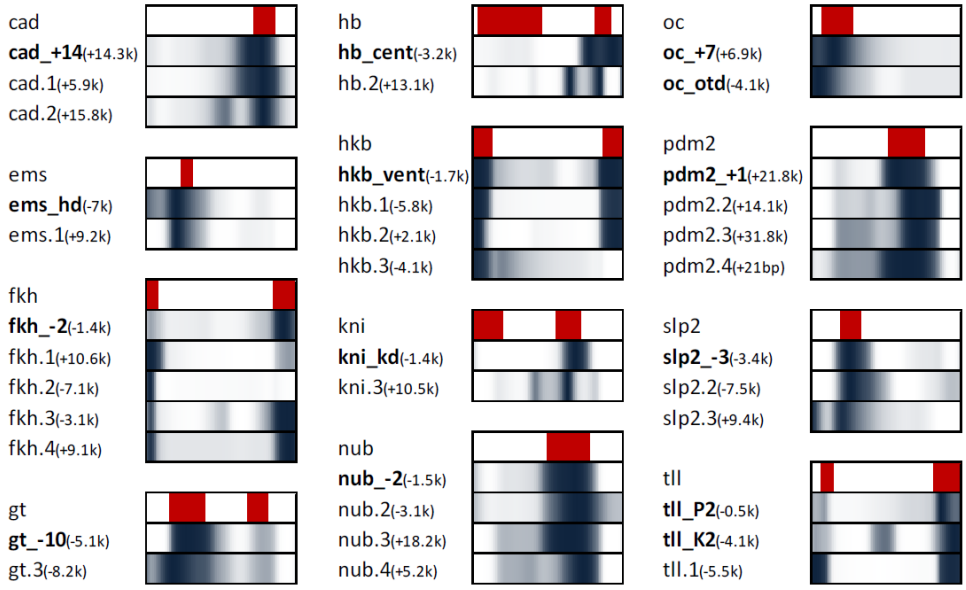


Figure 5.2 Redundant Putative Enhancers. Several of the 22 A/P genes have two or more related enhancers (either predicted or known) that drive similar expression patterns. For each gene, the endogenous gene expression is shown (red), along with predicted expression profiles of identified enhancers (blue). Labels in bold indicate known enhancers. Predicted expression pattern is shown with color intensity proportional to expression value.

We applied the PGP method to a larger collection of 144 genes with patterned expression along the anterior-posterior axis [146]. We automatically extracted the A/P expression profiles of these genes from the FlyExpress database [144], transformed the intensity values into binary expression domains, and identified flanking sequences with significant pattern generating potential at the same empirical p-value described above. Overall, we identified 123 putative enhancers from 68 genes, of which 44% overlapped a ChIP-chip peak (at 1% FDR; 65% when considering peaks at 25% FDR). The predictions included enhancers for genes with a single expression domain and genes with multiple expression domains (e.g., *slp1* and *ara*, respectively). Among enhancers corresponding to genes with multi-domain patterns, 53% capture only one of the domains of the endogenous pattern (e.g., *drm*); while 47% capture more than one domain (e.g., *emc*).

Sixteen of the above enhancer predictions overlapped previously verified regulatory sequences, of which 12 have blastoderm stage expression that agrees with the predicted expression profile from our model. These provide an independent experimental validation for our enhancer activity prediction pipeline. In addition, we tested seven enhancer predictions using new reporter transgenes. These lines were created as part of an ongoing project to systematically examine regulatory regions surrounding a subset of *Drosophila* genes with patterned expression in the nervous system [147]. Only predictions in genes with intergenic or intronic regions of at least 10 kbp were chosen for analysis. Selections included regions flanking genes with "strong" or "weak" A/P patterned expression. Four of 7 tested regions exhibited reporter gene expression patterns resembling the predicted pattern. For one of these, *Ubx*, reporter expression is in the correct region of the embryo, but initiation of the pattern is delayed relative to the endogenous gene. All three of the remaining tested reporters exhibit expression in the developing CNS, where many of the same TFs that regulate A/P patterning are expressed. It is possible that the same combinations of TFs that predict an A/P pattern in our model act to direct patterned expression in the developing CNS. We note that the specificity we observed here (57%) is about the same as that recorded in cross validation tests on the A/P gene set.

5.2.3 Construction of regulatory networks

Unlike other methods of enhancer discovery that rely on binding site clustering [34, 101], the PGP method incorporates both the binding specificities of TFs and their expression pattern to identify and predict the expression activity of an enhancer. Using the PGP method, it is possible to computationally assess the contribution of each TF to the enhancer by asking if altering the expression of the TF affects the quality of the prediction. We employed this strategy to infer direct regulatory interactions between TFs and enhancers, depicted as edges in the transcriptional regulatory network. To visualize the effect of removing an individual TF from the model, we simulated a "knock down" of the transcription factor (by setting its motif score to 0) and compared the predicted enhancer expression in this "*in silico* mutant" background and in "wild type". Unlike traditional *in vivo* genetic assays where observed changes may be the indirect effect of mis-regulation of other genes, this approach examines the direct contribution of a TF to a specific enhancer. In order to assign a statistical significance to this contribution, we created a null distribution of PGP similarity scores by generating random activity profiles from permutations of the TF's concentration profile and comparing them to the "true" activity. The

score obtained with the actual profile is compared to this distribution to produce an empirical p-value. When there are few binding sites in the enhancer, the TF pattern has little influence on enhancer predictions and the null distribution of scores is very narrow. When there are more binding sites in the enhancer, there is a broader distribution of similarity scores from the random profiles and the position of the actual profile within this distribution reflects the combined contribution of the binding sites and the normal TF expression pattern on enhancer activity. Using this procedure to infer a p-value for every TF-enhancer combination, we constructed a transcriptional regulatory network (involving the 35 enhancers where the model's quality of fit was not poor).

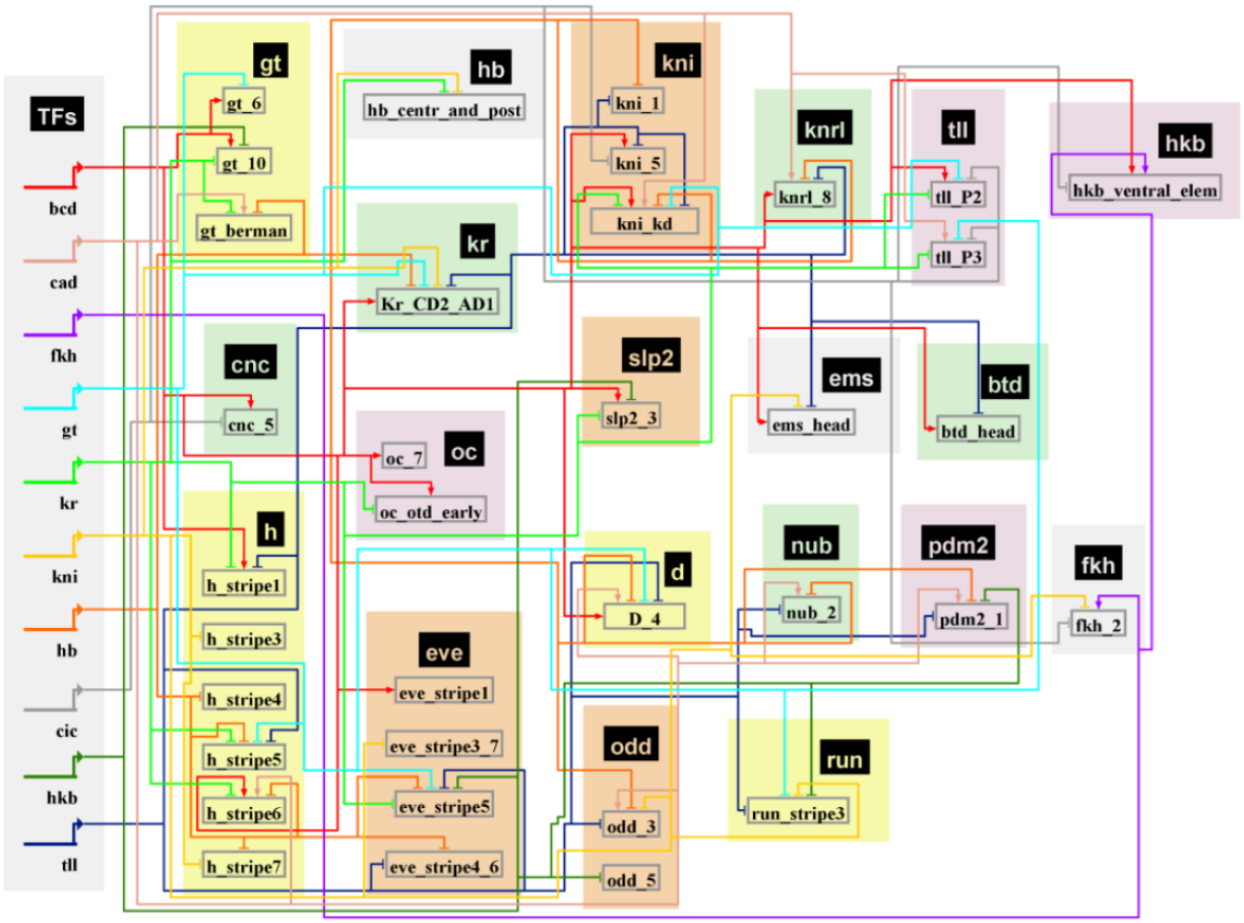


Figure 5.3 Inferred Regulatory Network Inferred Regulatory Network of 10 TFs and 35 enhancers. Colored edges indicate a regulatory influence between the corresponding TF on the left and the enhancers on the right grouped by their related gene (shaded box).

A total of 102 regulatory edges were predicted (at p-value < 0.05) between the 10 TFs and 35 enhancers, revealing a very dense network (Figure 5.3). 82 edges were supported by ChIP-based evidence of occupancy at the strongest level (1% FDR). 63 of the 102 edges have

been previously reported in the literature, mostly by examination of enhancer activity in mutant embryos lacking the TF. In some cases, confidence in experimentally determined TF-enhancer edges is further increased by *in vitro* confirmation of TF binding sites by DNaseI footprinting. For 12 of the 35 enhancers analyzed above, the FlyReg database [148] catalogs at least one such interaction with either BCD, CAD, KR, KNI, HB, GT, or TLL. These validated TF-enhancer edges were significantly enriched in our network (Hypergeometric test, p-value = 0.0026). Remarkably, the PGP-based regulatory network exhibited a greater enrichment for the validated TF-enhancer interactions than the ChIP-derived network, primarily by predicting fewer interactions with higher precision. Among the examples of interactions predicted by ChIP, but not PGP, we found multiple surprising examples of ChIP data indicating TF occupancy that should adversely affect the module's expression profile. Specifically, we identified enhancers with ChIP signals for the repressors KR, KNI, or GT, and whose activity domains overlap the bound repressor. Overall, we found 19 such cases of apparently "incongruous" occupancy. In 17 of these cases, we did not find corresponding support for evolutionarily conserved binding sites from multi-species motif profiles. These examples indicate a discrepancy between motif-based evidence and ChIP evidence, and suggest that the observed biochemical occupancy does not act to shape the activity pattern of the enhancer.

We applied the above statistical procedure to construct a regulatory network from all enhancer predictions (62 in the 22 A/P genes, 123 in the set of 144 genes). Analysis of the predicted network revealed several common patterns. A recurring theme in the TF-enhancer interactions was that of potential "auto-regulation" by activators. For example, all three predicted modules near the *cad* gene had significant regulatory input from CAD. In each case, this predicted auto-regulation was supported by ChIP data (at 1% FDR). Similarly, 4 out of 5 predicted modules for *fkh* are predicted to have FKH-driven activation. *Fkh* auto-regulation (in salivary glands) has been experimentally shown by [149]. On the other hand, auto-regulation by repressors is not seen in our predictions, as anticipated. Another common theme observed was that of mutual repression by pairs of TFs, e.g., HB – KNI, GT – KR, KR – KNI, HB – KR, GT – KNI, and TLL – KR, some of which were reported previously [138, 150-152]. We also characterized "complexity" of the enhancers with the edges of the network. Each enhancer on average had about three incoming regulatory TF edges, except enhancers driving expression in the anterior seem to have relatively low complexity.

5.3 Enhancer modeling in poorly characterized cell types

In this section, we show that similar enhancer activity models can be developed and applied to systems where validated enhancers with known expression patterns do not exist and where the set of relevant transcription factors are unknown. Specifically we will assess how well we are able to annotate regulatory elements that control gene expression in these poorly characterized cell types using TF motifs and limited experimental data.

5.3.1 Identifying putative enhancers and preliminary functional assignments

High throughput chromatin state (e.g., DNaseI hypersensitivity) data has been applied to identify putative enhancers in the genome [40, 48, 72, 130, 153-155]. However, these approaches typically do not associate enhancers with genes and expression domains. We sought to predict the target gene and expression domain of putative enhancers using enhancer activity models that incorporate the predicted TF motif profiles and TF-domain associations from Chapter 3.

We began by evaluating several types of genome-wide assays to identify the best method for locating putative enhancers, using 684 non-overlapping REDfly enhancers [59] as a benchmark. For each REDfly enhancer of length 100 to 3000 bp, we selected a size-matched, random intergenic region near an early development gene. We then calculated the values of several features for these 1,368 regions. We downloaded ChIP-chip datasets for the CREB Binding Protein (CBP) at 11 developmental time points and for 6 histone marks for 6 developmental time points from ModENCODE [17] and downloaded the *Drosophila* phastCons track [156] from UCSC Genome Browser [157]. We also examined the DNaseI chromatin accessibility from BDTNP and each of the 325 multi-species motif scores from Chapter 3. Multiple time points for CBP, histone marks, and accessibility were combined by selecting the best scoring time point per region. The 325 motif scores were summarized by the best per region, the average of all, and the number that were significant. These score features are independently evaluated for their ability to identify REDfly enhancers by the AUROC metric.

Open chromatin, as indicated by high accessibility scores, was found to be the best method with an AUROC of 0.789. The occupancy profiles of the general transcriptional co-activator CREB Binding Protein (CBP), as well as histone marks associated with enhancer and promoter regions (H3K4Me3, H3K4Me1, H3K9Ac, and H3K27Ac) were also predictive, while phastCons scores of evolutionary conservation and methods based on combining motif scores were considerably worse at discriminating REDfly enhancers (Figure 5.4).

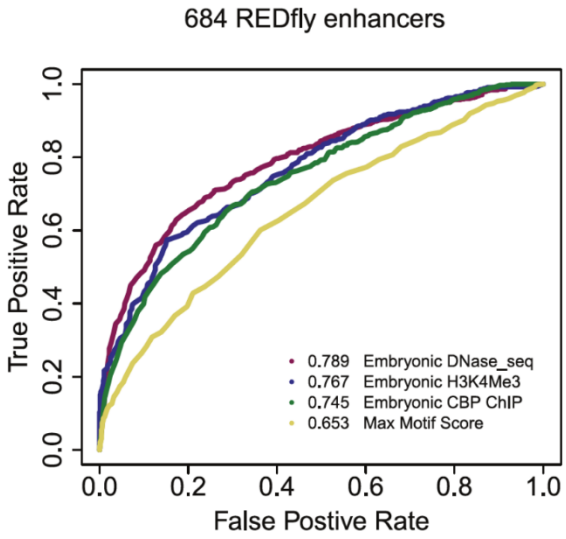


Figure 5.4 Discriminative Features of REDfly Enhancers. The ROC curves for methods of detecting 684 REDfly enhancers from 684 negative sequences. AUROC for each method is reported in the legend.

These observations confirmed our decision to define our set of putative enhancers as those non-overlapping 500 bp segments that are among the top 10% most accessible regions in any of the four developmental stages: 5, 9, 11, and 14. In Chapter 3, we also demonstrated the utility of this single experimental assay in combination with motif scoring profiles in identifying major regulators of cell type specific co-expressed gene sets. We did not consider any genomic segments that overlapped exons or regions of tandem repeats by more than 50% of their length. As an additional filter, we only considered segments whose combined multi-species Brownian Motion motif score (sum over all 325 motifs) was above a threshold of 10. This filter was motivated by observations of the summed motif score distributions of REDfly enhancers. We henceforth refer to this putative set of enhancers defined by these accessible genomic segments as "open regions". Table 5.2 shows that these regions are highly enriched for gene-proximal locations ($\leq$ 5 kbp upstream of transcription start sites), similar to [106].

Location	OpenRegion%	Genome%	Fold Change
FarUp	2.2%	5.2%	0.43
MedUp	5.6%	6.8%	0.83
NearUp	21.5%	10.2%	2.11
CDS	0.0%	13.3%	0.00
Intronic	53.4%	45.1%	1.18
NearDown	10.7%	8.8%	1.21
MedDown	4.8%	6.2%	0.77
FarDown	1.8%	4.5%	0.39

Table 5.2 Distribution of Open Regions. Regions are assigned a label based on their position relative to their nearest gene depending on whether they are intronic, exonic (within CDS), or >20 kbp (Far), 5-20 kbp (Med), <5

kbp (Near) from the upstream or downstream end of the gene. The distribution of the open regions (in first column) is compared to the genome-wide distribution (in second column) with their fold change (in third column).

5.3.2 Model training and evaluation metrics

We sought to annotate the collection of open region for their most likely activity in the 195 BDGP [113] based expression domains we defined in Chapter 3. We created a preliminary, "noisy" assignment of each enhancer to one or more expression domains based on gene proximity, gene expression annotations, and the accessibility profiles of enhancers. For a given expression domain, *D*, all open regions accessible during the appropriate developmental stage with at least one within 5 kbp neighboring gene annotated with *D* were preliminarily assigned with that expression activity. On average, about 14 expression domains were tentatively assigned to each enhancer, suggesting that further methods are required to resolve ambiguities.

To further refine these tentative domain assignments, we learned computational models (classifiers) capable of predicting expression driven by an enhancer. This requires training sets of "positive" and "negative" examples, i.e., open regions known to drive or not drive expression in a particular domain. Reliable training sets of this type are rare for most expression domains. Enhancers from the REDfly database may be used for training models, but this would limit the model training to relatively few expression domains. Instead, we chose to train models on the numerous open regions putatively assigned to each domain, so that the positive (negative) training sets are likely to be enriched in (depleted of) enhancers of an expression domain. Incorporation of these "noisy training sets" also allowed us later to treat REDfly enhancers as "unseen" test data for evaluating the models.

For each expression domain, *D*, we selected up to 500 "noisy" positive enhancers and an equal number of negative enhancers. We marked ¾ of these data for training and the remaining ¼ for testing. The model is trained and the "test" AUROC is recorded. For the 40 expression domains with at least 10 open regions overlapping REDfly enhancers annotated with the domain, additional testing sets "REDfly vs. Open Regions" (RFVO) and "REDfly vs. Enhancers" (RFVE) are created. Both RFVO and RFVE test sets take the open regions overlapping the REDfly enhancers as positives. The negatives for the RFVO test set are sampled from the negatives from the general testing set. For the RFVE set, the negatives are chosen from open regions that are accessible during the developmental stage of *D* and overlap REDfly enhancers that are not annotated *D*. The RFVO test set allows us to evaluate if our classifier distinguishes

enhancers of the expression domain from other open regions, while the RFVE test set evaluates our ability to distinguish enhancers of the expression domain from other validated enhancers of other expression domains. The AUROC for the RFVO and RFVE test sets for the 40 applicable expression domains are also calculated and recorded.

5.3.3 Formulation of enhancer activity model

For each expression domain, we trained a "complete" linear model to discriminate positive and negative open region examples using features that correspond to each of the 325 TFs in our collection and each of the four stages of development. Each TF-related feature was the product of four quantities: the multi-species motif score of the TF in the open region, the strength of statistical association between the TF's motif and the expression domain, the expression annotation of the TF's gene in the given expression domain, and the RNA-seq expression level of the TF's gene in the appropriate developmental stage. Accessibility scores of the open region in each of the four developmental stages were also included as features modeling the open region.

The activity-prediction model (henceforth called the "complete" enhancer model) for a domain *D* is formally described as:

$$y^r = \sum_{m=1}^{325} \alpha_m Z_m^r S_m^D E_m^D R_m^D + \sum_{s=1}^{4} \gamma_s A_s^r + \beta$$

where

- y^r is the prediction indicating whether region *r* is in the positive set
- *m* is one of the 325 motifs
- *s* is one of the four developmental time points (stage 5, 9, 11, and 13)
- α_m, γ_s, and β are the domain-specific parameters
- Z_m^r is the non-negative multi-species motif scores for region *r* for the m^{th} motif
- S_m^D is the negative logarithm of the p-value of association between the expression domain *D* and the TF represented by the m^{th} motif
- E_m^D indicates whether the TF related to the *m*th motif is expressed in *D* or in a related expression domain

- R_m^D is the "fragments per kilobase of exon per million fragments mapped" (FPKM) reported from [158] for the TF related to the m^{th} motif in the developmental stage related to expression domain *D*
- A_s^r is the chromatin accessibility score for region *r* for the s^{th} developmental stage.

In Figure 5.5, we illustrate the training set selection, the model features and formulation, and our procedure annotating cell type specific enhancers (explained in the Section 5.3.5).

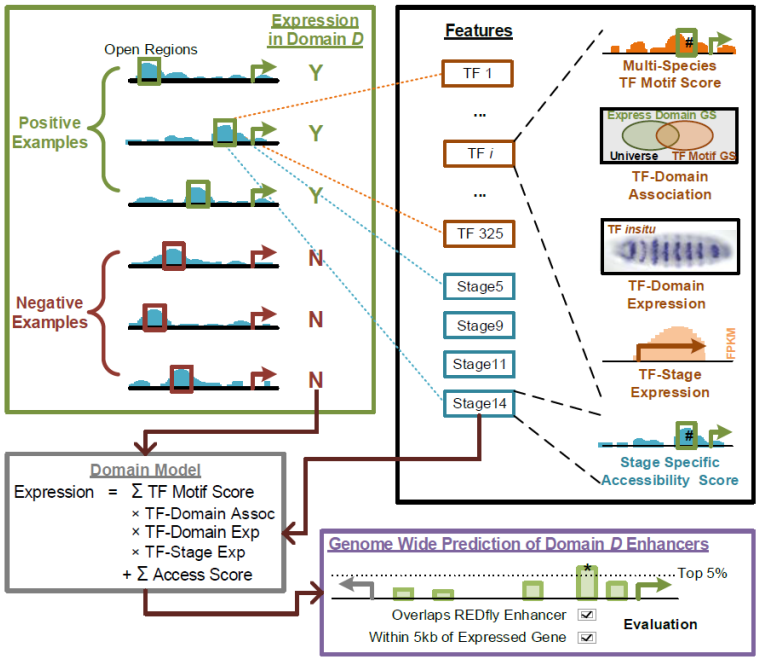


Figure 5.5 Enhancer Modeling Pipeline. To train an expression domain specific model of enhancer activity, our linear model combines each putative enhancer's accessibility features with TF features that are the product of the motif score, the importance from TF-domain analysis, and the TF's expression from *in situ* annotations and from RNA-seq data. "Good" models (RFVO AUROC > 0.7 or test AUROC > 0.6) are applied to annotate likely enhancers of domain specific expression genome-wide.

Our "complete" linear models exhibited an AUROC of at least 0.7 on RFVO test sets from 21 of the 40 expression domains where RFVO evaluation was possible. Sixteen of 40 linear model classifiers exhibited an AUROC of at least 0.7 when evaluated with the RFVE test set, which did not occur in any negative controls. For the remaining 155 expression domains, REDfly evaluations were not possible and AUROCs were obtained using "left-out" test sets from the noisy training sets. Fifty-six of these expression domains exhibited a test AUROC of at least 0.6, a level of discrimination observed on only 3 of 155 domains in negative controls. We allow for the lower AUROC threshold because we expect the model will be unable to capture these noisy test sets as well. Overall, we learned accurate models for 77 of the 195 expression domains.

5.3.4 Comparison to other models

We applied the same evaluation framework to compare the "complete" model (containing TF motif, TF expression, and accessibility information) to simpler variants that ignored certain types of features. For instance, we found the complete model to accurately predict more expression domains than analogous linear models that incorporate only motif features ("Motif * Express") or only accessibility features ("Access") (Figure 5.6). The advantage of using motif features over only accessibility-based features was most conspicuous for earlier expression domains prior to developmental stage 13.

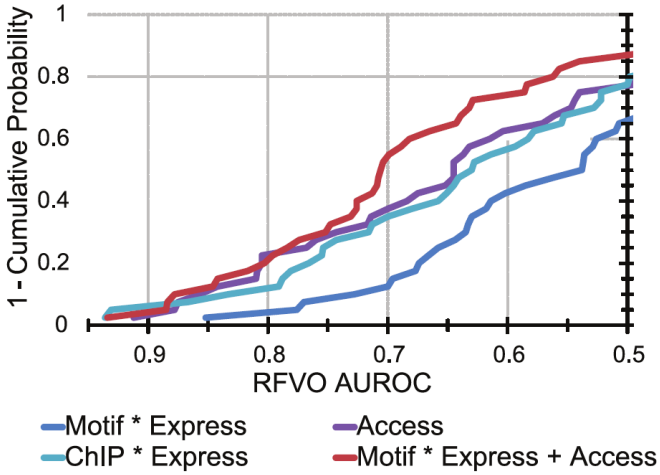


Figure 5.6 Comparison of RFVO AUROCs. For four different model constructions, we calculated the corresponding AUROC using the RFVO test set on each of the 40 expression domains. The distribution of these forty values is visualized with the x-axis showing a particular value of the AUROC and the y-axis indicating the percentage of the domains with a stronger AUROC. Of the four models compared, the best model, "Motif * Express + Access", combines 325 motif based features with four accessibility based features in a linear model (see panel C).

Since our approach incorporates computationally predicted TF-DNA binding, it is reasonable to compare it to a baseline that utilizes TF-DNA binding data from ChIP experiments in a similar manner. To this end, we trained an alternative classifier where TF-related features utilized 69 publicly available genome-wide ChIP profiles rather than the 325 motif profiles computed by us (see Chapter 3).

The form of the full enhancer model using ChIP and TF expression data for a domain *D* is described as:

$$y^r = \sum_{c=1}^{69} \alpha_c \, C_c^r \, E_c^D \, R_c^D + \beta$$

where

- y^r is prediction indicating whether region *r* is in the positive set

- c is from the 69 ChIP datasets
- and α_c and β are the domain-specific parameters
- C_c^r is the averaged ChIP score for the region r for the cth ChIP set
- E_c^D indicates whether the TF related to the cth ChIP dataset is expressed in D or in a related expression domain
- R_c^D is the fragments per kilobase of exon per million fragments mapped (FPKM) reported from [158] for the TF related to the cth ChIP dataset in the developmental stage related to expression domain D.

The "complete" motif-based model performed accurately on more expression domains than the ChIP-based models (Figure 5.6), suggesting that having computationally characterized TF-DNA binding features spanning more TFs is better than relying on experimentally characterized occupancy for fewer TFs. On closer examination, we noted that an improved performance of motif-based models over ChIP-based models frequently corresponded to expression domains from developmental stages 13-16. This may be because of poor temporal resolution of these stages in the available ChIP data or because the crucial TFs of these later stages have not yet been subjected to ChIP assays.

We also compared our linear classification method to other classification schemes such as logistic regression and support vector machines. Overall, we found the linear model to perform marginally better. The regression model also has the advantage that the explicit activity pattern predictions are easily interpreted, compared to other machine-learning techniques such as Bayesian networks [49] or support vector machines [38].

5.3.5 Application to enhancer annotation

We next attempted to assign expression activity to putative enhancers using the motif-based models trained as above, focusing on the 77 expression domains for which such models were assessed to be accurate. We attributed an expression domain to an open region if one of the neighboring genes is annotated with the domain and the complete model for the domain scored the open region in the top 5% of all 23,529 open regions genome-wide (Figure 5.5). This resulted in a compendium of 7,824 high-confidence enhancer activity predictions spanning 4,197 open regions. Over 30% (2,354) of these predictions involved putative enhancers located > 5 kbp away from the target gene. A large number corresponded to annotated REDfly enhancers, even though these enhancers had not been seen in training models. In order to evaluate the accuracy of

genome-wide enhancer activity prediction for each REDfly enhancer, we examined the strength of its association with each possible expression domain (as predicted by the appropriate model) and found that the experimentally annotated expression domain ranked first significantly more often than expected by chance. This result was stronger with predictions by the motif-based models than with equivalent predictions by ChIP-based models. One successful example of enhancer activity assignment procedure comes from the *string* (STG) gene locus (Figure 5.7). In this region, there are a number of REDfly enhancers annotated to drive expression in the ventral nerve cord and the ventral epidermis. We highlight five open regions in this locus whose predictions for domain specific expression agree with the known expression patterns of overlapping REDfly enhancers.

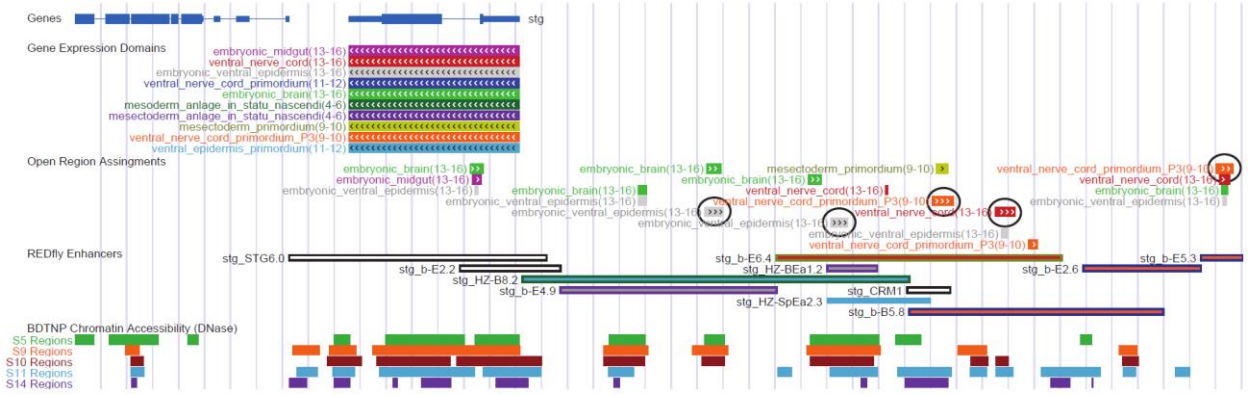


Figure 5.7 Annotated Enhancer Example. Genome browser view of enhancer predictions near stg gene with the position and structure of genes is shown at the top. At the bottom, the chromatin accessibility from DNaseI-seq of four developmental time points is shown as colored profiles. Each possible expression domain of stg is shown ("Gene Expression Domains") and color-coded. The "REDfly enhancers" are shown with the fill and border color matching their annotated gene expression domains. Finally, the "Open Region Assignments" show which expression domains are likely driven by each 500 bp open region. The color and size of the open region box indicate the driven expression domain and the significance of the prediction. Five different open regions are circled where the most significant expression domain prediction is consistent with the annotation of an overlapping REDfly enhancer.

5.4 Discussion

In this chapter, we presented enhancer modeling approaches to address the problem of annotating enhancers throughout the genome. We began by creating a model of enhancer activity in the well-studied A/P segmentation system from validated collections of known enhancers and TF regulators. We demonstrated how to annotate additional enhancers on A/P genes from this model and to build regulatory gene networks that were more specific than ChIP. We then attempted to assess this enhancer modeling approach on 195 cell types in the *Drosophila* embryo, most of which have no validated enhancers and very limited knowledge of TF regulators. With the incorporation of stage specific chromatin accessibility information and our

methods for identifying regulatory TFs, we were able to build and apply models of enhancer activity for over 75 of the cell types.

In the study of the A/P system, we introduced a novel similarity score for comparing a predicted enhancer activity pattern to the expression of its neighboring gene. We applied this pattern generating potential (PGP) as a tool to annotate the non-coding genome. Unlike the similar "regulatory potential" score of [159], which generally classifies non-coding sequences as regulatory or neutral, PGP scores sequences by their ability to contribute to the specific expression pattern of a nearby gene. It further facilitates a quantitative inference of TF-enhancer interactions, whose validity may then be assessed though *in vivo* observations. We have specifically applied this approach to the A/P network, but it should be applicable to any system in which adequate expression data is available for relevant TFs, enhancers, and target genes. It is especially relevant to systems with complex expression patterns that may include distinct spatial and temporal dimensions. One example application is suggested by [160]. In their work, they categorized lateral gene expression in the *Drosophila* oocyte as unions and intersections of two-dimensional primitive patterns. These primitive patterns are mapped onto the oocyte in order to partition it into non-overlapping regions, which may form the bins of expression in the PGP modeling framework. There are many other datasets where these techniques could be applied where complex gene expression patterns are captured by automated image-processing pipelines.

The logistic and linear models presented in this chapter are "simpler" than thermodynamic models of the sequence to gene expression relationship [44, 45]. At the same time, they perform well compared to the thermodynamic model and have the added advantages of easily incorporating multiple species comparisons and of computations that are orders of magnitude faster. This enables fast, genome-wide prediction of other enhancers, examination of the effect of each motif on each putative enhancer, and empirical assessment of its statistical significance through permutation tests. However, the regression model does not incorporate known mechanistic features of enhancer function, such as cooperative TF binding. More detailed models of enhancer function have been developed for individual enhancers [151, 161, 162], which accurately describe changes in enhancer activity over developmental time or due to mutation. While models with additional parameters may provide better predictions, they also require additional prior knowledge and may not generalize as well. We also note that the TF

motif scoring incorporated in our models are based on evolutionary conservation at the ~500 bp resolution and are thus likely robust to local turnover of sites [163].

Finally with high throughput technologies becoming the norm [164] for predicting enhancer locations, the challenge of enhancer functional annotation is increasingly important. Our work represents one of the most ambitious attempts to date at tackling this challenge, assigning activity to enhancers for as many as 77 of the 195 expression domains. Prior work in the field has attempted this with one [44, 45] or a handful [38, 39, 48, 134] of domains. Since validated training datasets are generally not available for most tissues, we considered the possibility of defining "noisy" training sets of enhancers active in an expression domain based on their accessibility and the distance and expression of their nearby gene. This pragmatic choice allowed us to successfully build regulatory maps for many domains beyond the handful with validated enhancers. We found our motif-based approach to annotate enhancer activity to be as effective as an analogous approach based on ChIP data. This is not a fair comparison since one method incorporates motifs for 325 TFs and the other relies on ChIP data for 40 TFs. However, the comparison should be interpreted in light of the costs of generating equivalent data for the two methods, a single accessibility profile for the domain versus hundreds of ChIP-seq experiments.

6 CHARACTERIZING GENE SETS WITH RANDOM WALKS ON HETEROGENEOUS BIOLOGICAL NETWORKS

6.1 Background

In the previous chapters, we have focused on characterizing co-expressed gene sets by common transcriptional regulatory features of the genes. However, this is only one way in which the co-expressed genes might be related; these genes may also exhibit other relationship such as shared protein domains, evolutionary origins, biological processes, etc. The experimental techniques to characterize genes and the public databases of curated annotations are rapidly increasing and incredibly diverse. There are resources with data on gene sequence conservation (e.g. OrthoDB [165]), protein sequence function annotation (e.g. Pfam [166]), condition specific transcript expression levels (e.g. GEO [167]), physical and genetic protein interactions (e.g. BIND [168]), associations of genes with diseases (e.g. OMIM [169]), detailed reaction pathways (e.g. KEGG [11]), curated annotations of proteins of their cellular localization and function (e.g. Gene Ontology [10]), binding to and chemical marks of chromatin (e.g. ENCODE [86]), etc. This chapter will address the challenge of incorporating these heterogeneous data from multiple sources ("types") into the task of characterizing a given gene set and identifying additional genes that are important and related.

One broad approach that researchers employ to perform analysis with these different public resources is to represent the data in a biological network. Rather than using each data source, one at a time, to analyze a co-expressed gene set, sources may be integrated within a network and simultaneously leveraged to identify related genes. This idea was rigorously tested in the MouseFunc challenge [52] where nine algorithms for integrating genomic evidence in a heterogeneous biological network of mouse genes were evaluated for their ability to discover genes functionally related to a given gene set. Other network-based gene ranking algorithms have been applied to the important tasks of identifying driver genes in cancer [53], potential gene targets for drugs [55] or microRNAs [170], and disrupted protein complexes involved in disease [171]. Network-based analyses of gene sets have also been designed to extend and annotate gene modules [172], quantify gene set enrichment for functional molecular networks [51], identify

frequent subnetworks shared across multiple diseases [173], or cluster and find signatures of cancer subtypes [54, 174].

Most gene set analyses performed on a biological network from heterogeneous data sources discard a majority of the data in the construction of the network. Most commonly [51, 53, 54], the rich and diverse public datasets are converted to homogeneous gene-gene networks: these vastly simplified networks contain only nodes representing genes of a single species and unweighted edges of a single type. In these homogenous networks, the edges only represent a relationship between a pair of genes, but details about the number, types, and strength of the evidences for that relationship is lost. Algorithms that rely on these networks assume that all relationships in the network are as reliable as any other. Others improve upon this unweighted, homogeneous (one edge type) network of gene-gene interactions by weighting the edges based on the strength of relationship (e.g. transformed correlation values [51]). However, the calculation of these weights often involves the assumption that each type of relationship (i.e. source database or experimental assay) is equally valuable. There are several papers with specific applications (e.g. identifying interactions between pharmacological drugs and their protein targets [55] or between genes and diseases [56]) that integrate biological networks containing more than one edge or node type. However, the networks in these papers usually have a structure specific to their system of interest; most often containing nodes of two different types and three types of edges capturing similarity within each type of node sets and the known relationships between them. Although they construct heterogeneous networks, they strictly rely on the structure of the problem and do not attempt to incorporate data from all possible sources.

GeneMANIA [57] is a popular, network-based gene ranking algorithm that performed well in the MouseFunc evaluations. The GeneMANIA approach specifically integrates data from many different sources without sacrificing the edge source information. Data from each source informs the creation of its own “affinity” network of gene-gene interactions. The multiple affinity networks are up- or down- weighted based on their relevance to the original functional gene set before being combined into a single composite network [175]. While the GeneMANIA approach works well and specifies the types of edges that are most important to the ranking task, it still discards the specific details about the gene-gene relationship when constructing each affinity network. For example, the edges within a Gene Ontology affinity network indicate that a

pair of genes share a GO annotation, but does not preserve which annotation(s) that it may have been.

Our goal in this chapter was to develop an algorithm that identifies genes related to a given set using biological networks that maintain detailed information from public data sources. Our algorithm was explicitly designed to work on networks with heterogeneous node and edge types that represent the complete collection of public knowledge. We relied on the algorithm to perform the gene ranking task and simultaneously return the specific, relevant network features. Like many other network ranking algorithms that rely on guilt-by-association approaches [53, 54, 176], our algorithm implemented a modified random walk with restart (RWR). However, unlike other methods, we employed a first round of RWR to simplify our large, noisy network of all public data and to report the features related to the given gene set. We found improved ranking results after a second stage RWR using only the relevant features of the original network. We evaluated our method's ability to recover left out genes from the expression domain gene sets of *Drosophila* embryonic development from Chapter 3 and Chapter 5. We showed that our gene ranking method improves when multiple data sources are combined and when additional species are added to create the original network. We finally applied the algorithm to a multi-species study of aggression in social animals [7].

6.2 Building a heterogeneous network

Our first task was to construct a heterogeneous network, which represents specific information from multiple public resources. We started by adding a "gene" node to our network for every gene in a species. We connected a pair of gene nodes with an undirected "homology" edge with significant protein sequence similarity if their BLAST e-value score [177] was less than 0.01. Additionally, we assigned weights to the homology-based edges that are calculated from the z-transform of their e-value significance (maximum value is set to a z-score of 8). We then created "feature" nodes in the network that represent computationally or experimentally derived characteristics of genes. The feature nodes derived from the same data source are said to have the same "feature type". A feature node was always connected to genes nodes by undirected edges of the same feature type with weights proportional to the reliability of the feature annotation. To incorporate protein structure data into our network, we first created ~3,700 new feature nodes of type "prot_domain", each representing a protein domain from Pfam [166]. We then connected each "prot_domain" type feature node to all of the gene nodes whose protein

contained that domain, as identified by HMMER [178] scans. The weight of the new edge was the thresholded z-transform of the HMMER e-value score of that domain in that gene. Homology and protein domain information was included for every species included in the network.

Additionally, in our *Drosophila melanogaster* network, we incorporated hundreds of feature nodes of type "motif" that represent distinct TF binding specificities (motifs). A motif node connected to the genes whose 5 kbp upstream regulatory region contain the motif, i.e., if the regulatory region includes one of the top 0.5% of the highest scoring 500 bp windows genome-wide as scored by the Stubb program for that motif [35]. The weights on these edges were the z-transform of that window's empirical p-value (see Chapter 3). Also for the *D. melanogaster* network, we incorporated feature nodes of type "ChIP", representing TF occupancy obtained from each of 75 ChIP-seq experimental datasets corresponding to the early fruit fly embryo (see Chapter 3). Each "ChIP" type feature node represented an experimental assay and was connected to a gene if the TF in the developmental stage assayed binds to the gene's 5 kbp upstream gene regulatory region. The specifics of assigning weights to these edges were the same as those for "motif"-gene edges.

For the network used to study aggression across species, we defined 1,827 feature nodes of type "Gene Ontology", each one representing a term from Gene Ontology [10]. GO annotations for three species (human, mouse, and fly) were downloaded from Ensembl [179] and only terms with at least 20 annotated genes across the three species became feature nodes and were connect to their annotated genes in the three species. The edges of this feature type had weight 2 if the corresponding GO annotation was curated and weight 1 if it was inferred computationally. Also for the aggression study, we added 12 mouse-specific "brain atlas" feature nodes derived from gene expression information produced as part of the Allen Brain Atlas [4]. Each feature nodes corresponded to a specific region of the mouse brain and each connected with an edge of weight 1 to the 100 genes that are most specifically expressed in that region.

For each application of our algorithm, we created a weighted, undirected network, choosing some or all of the above-mentioned components, as appropriate. Given k selected feature types, our initial network was constructed with gene nodes *G* and sets of feature nodes for each different type, $F_1, F_2, \ldots F_k$, (e.g. "motif", "brain atlas", etc.). We represented the edges of this network with an adjacency matrix with the form

$$M = \begin{bmatrix} M_{GG} & M_{GF_1} & \cdots & M_{GF_k} \\ M_{F_1G} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ M_{F_kG} & \cdots & \cdots & M_{F_kF_k} \end{bmatrix}$$

where all of the homology type edges were contained in the submatrix M_{GG}, while M_{F_iG} and M_{GF_i} were the submatrices that represent edges between all feature nodes of type *i* and genes in *G*. There were no edges between feature nodes, meaning $M_{F_iF_j} = \mathbf{0}$ for all *i*, *j*.

6.3 Functional annotation from two stage random walk

Given a biological network *M*, a novel experimental gene set (referred to as the "query" set *Q*), and the universe *U* of genes to rank, we employed a two stage algorithm based on a modified random walk with restart (RWR) approach [180] to rank the gene nodes of *U*. The algorithm also ranks the feature nodes in the network by their relevance to the query set *Q*. One may understand the effect of a RWR algorithm by imagining a walker on a node in the network. With probability (1-*c*), where *c* is the restart parameter, the walker follows an outgoing edge to a neighboring node and with probability *c*, the walker resets to one of the genes in the "restart set", defined as the query set *Q* in our algorithm. In properly formed networks over the long run, the probability distribution of the walker over all nodes will converge to a so-called stationary distribution. This distribution produces a ranking on all nodes that incorporates the connectedness of the node in the network as well as the proximity of the node to the query set. In the first stage of our algorithm, we applied RWR to find the highest ranking feature nodes to extract a relevant subnetwork of the initial network. The results of the second stage RWR on the subnetwork provide us the final rankings of nodes in *U*. Both stages are described in detail below.

6.3.1 Algorithm design

Before applying our RWR algorithm, we first must normalize the edge weights in the initial heterogeneous, biological network. We first normalized the weights of all edges of the same type (e.g. all homology edges, or all edges connecting genes to nodes of type "prot_domain") to create the normalized adjacency matrix *N*. In terms of our notation, for any two sets of nodes of a given type (X and Y) where at least one is the set of gene type nodes:

$$(N_{XY})_{i,j} = {(M_{XY})_{i,j}}\Big/{\sum_{i,j}(M_{XY})_{i,j}}$$

We did this to equalize the global probability of the walker following a specific edge type. For example, even though "motif" type edges might account for 10 times the weight as the "prot_domain" edges, this heuristic adjusted the edge weights so the walker takes "motif" edges as often as "prot_domain" edges overall.

Next we normalized each of the columns the matrix *N* to form a transition matrix, *A*.

$$A_{i,j} = {N_{i,j}}\Big/{\sum_i N_{i,j}}$$

The value $A_{i,j}$ is the probability that the walker following an outgoing edge will transition from node *j* to node *i*.

We define $\boldsymbol{v^t}$ to be calculated probability distribution (or "relevance vector") of the walker over all nodes in the network after *t* steps of the RWR algorithm. We initialized this probability distribution, $\mathbf{v^0}$, to be the uniform distribution over all nodes by default. Each step of the random walk is notated as:

$$\mathbf{v}^{\mathrm{t+1}} = (1 - \mathrm{c})\mathbf{A}\mathbf{v}^{\mathbf{t}} + \mathrm{c}\boldsymbol{\alpha}$$

where *c* is the restart probability and $\boldsymbol{\alpha}$ reflects the probability of jumping to a gene in the restart set. When the restart set is defined as the set of query genes *Q*, then

$$\boldsymbol{\alpha}_i^Q = \begin{cases} {1}/{|Q|} & \text{for gene nodes in } Q \\ 0 & \end{cases}$$

As the random walk is irreducible and aperiodic, the iterative update of this procedure is guaranteed to converge to the stationary distribution of the random walk regardless of the initial probability distribution $\mathbf{v^0}$. We ran iterations of the RWR with the query set defining the restart set ($\boldsymbol{\alpha} = \boldsymbol{\alpha}^Q$) until the relevance vector converged ($|\mathbf{v}^{\mathrm{t+1}} - \mathbf{v}^{\mathbf{t}}| < 0.05$). We notate this converged probability distribution as $\tilde{\mathbf{v}}_{\boldsymbol{Q}}$. The ranking of all nodes by the probabilities of $\tilde{\mathbf{v}}_{\boldsymbol{Q}}$ is referred to as the "stage 1 query ranking". We repeated the RWR procedure using the set *U* of all genes we are trying to rank as the restart set (in place of set *Q* above). We arrived at a second converged relevance vector $\tilde{\mathbf{v}}_{\boldsymbol{U}}$ and refer to the ranking it induces on all nodes as the "baseline ranking". Note, $\tilde{\mathbf{v}}_{\boldsymbol{U}}$ captures the overall relevance/importance of each node in the network without consideration of the query set, whereas $\tilde{\mathbf{v}}_{\boldsymbol{Q}}$ incorporates overall network structure as well as proximity to the query set. Therefore, to find the feature nodes most specifically relevant to the query genes, we examine the difference between these vectors, $\tilde{\mathbf{v}}_{\boldsymbol{Q}} - \tilde{\mathbf{v}}_{\boldsymbol{U}}$, as described next.

For the second stage of our two stage RWR, we selected the 50k most query specific feature nodes (greatest values in $\tilde{\mathbf{v}}_{\boldsymbol{Q}} - \tilde{\mathbf{v}}_{\boldsymbol{U}}$,) and created a subnetwork *M'* from the initial matrix M by removing all non-selected feature nodes and their connected edges. Thus,

$$M' = \begin{bmatrix} M_{GG} & M_{GF_1'} & \cdots & M_{GF_k'} \\ M_{F_1'G} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ M_{F_k'G} & \cdots & \cdots & M_{F_k'F_k'} \end{bmatrix}$$

where F_i' represent only selected feature nodes of feature type *i.* Using the same normalization procedure as above, renormalized *M'* by type and converted it to the transition matrix *A*'. We repeated the random walk using *A*' and $\boldsymbol{\alpha}^{Q}$ (restart set defined from the query set *Q*) until we converged to the new relevance vector $\tilde{\mathbf{v}}_Q'$. The ranking of all nodes induced by this relevance vector was called the "stage 2 query rankings".

6.3.2 Evaluation of two stage RWR algorithm

We employed a cross validation scheme to evaluate the results of our ranking method. For each given gene query set, we held out 10% of the genes for testing, Q_{Te}, and the remaining 90% of the gene set are supplied to the algorithm as the query set Q_{Tr}. With a query set Q_{Tr} , we produced the "stage 1 query rankings", identified the relevant features nodes and extracted the query specific subnetwork, and repeated the RWR to produce the stage 2 query ranking. From the calculated rankings and the held out test sets Q_{Te}, we produced receiver operating characteristic (ROC) curves and quantified the performance of our algorithm with the area under these curves (AUROC).

6.4 Evaluations in *Drosophila* developmental cell types

We first applied our gene ranking and feature selection algorithm to the sets of genes defined from *insitu* images of gene expression in *Drosophila* embryos from BDGP [113]. For this analysis, we focused on 92 spatio-temporal expression domains that contained between 100 and 1200 genes with the specific expression pattern. We applied the algorithm to each expression domain gene set separately and evaluated gene rankings with the AUROC on the held out test set. In this application, we tested the feasibility of our algorithm to find additional genes related to each query set (using the AUROC measures described above). This application is important in instances where experimental annotation of genes has a non-trivial cost (as with image in situ

hybridizations). Finding related genes through our gene ranking procedure provides investigators a limited number of additional genes to assay.

We began by creating a *Drosophila* -specific heterogeneous network that contained the "homology", "prot_domain", "motif", and "ChIP" type edges described in Section 6.2. The number of nodes and edges of each feature type are described in Table 6.1. For each of the 92 expression domain gene sets, we ranked the 13,609 gene nodes in this network and reported on where the held out genes fall in this ranking. We also found the most relevant features nodes for each gene set.

FeatureType	nNodes	nEdges
homology	13,609	270,125
motif	223	222,191
prot_domain	3,579	34,244
ChIP	75	53,710
Total	**17,486**	**580,270**

Table 6.1 Composition of Drosophila Network. Lists the number of nodes and edges of each type in the heterogeneous network containing only *Drosophila* genes.

6.4.1 Results on *Drosophila* networks

The AUROC values on the 92 expression domain gene sets are shown in Figure 6.1. We observed that the rankings produced by our second stage RWR are better than the rankings from the first stage. For instance, the AUROC of the two stage procedure is > 0.6 for 76 of the 92 gene sets, while that of the first stage along is >0.6 for only 66 gene sets. (0.656 > 0.643). The improvement in the second stage presumably resulted from removing unrelated features for a particular query gene set from the random walk. Since we do not know *a priori* which features may be important to any given set, this two-stage approach allows us to begin with all known data encoded in the network, reduce to a relevant subnetwork, and produce better rankings. This is an important improvement over a majority of RWR algorithms that only produce rankings from the original networks that contain edges potentially irrelevant to the query gene set.

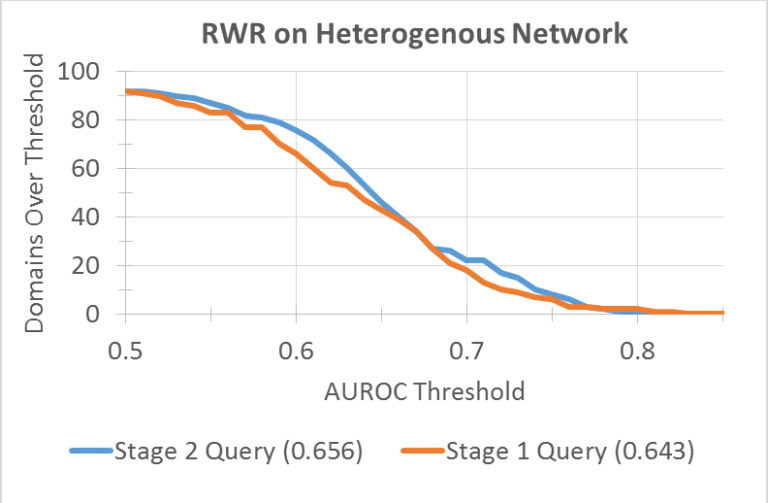


Figure 6.1 Comparison of Stage 1 and 2 Rankings on Drosophila Heterogeneous Network. We compared the rankings produced at the end of the first stage random walk to the second stage random walk on query specific networks. We calculated the average stage 1 and stage 2 AUROCs for each of the 92 expression domains and then plot the number of domains (y-axis) that were above each possible AUROC threshold (x-axis).

The next observation was that rankings are better due to our use of a heterogeneous network that combined data from multiple sources. Instead of the heterogamous network (with four different edge types) that was used in the test reported above, we produced four separate networks each with only the edges of a single type. We ran our two-stage algorithm on the 92 expression domain gene sets on each network and found that the heterogeneous network provides the highest AUROC on average (0.656). In general, the heterogeneous network outperformed the homogeneous "prot_domain" and homogenous "ChIP" networks, which were much better than the homogenous "motif" network (Figure 6.2). At high AUROC thresholds (0.8), the homogenous "prot_domain" network was able to correctly rank genes of expression domains including sensory system (stage 13-16), germ cell (stage 9-10), procephalic ectoderm AISN (stage 4-6), and embryonic anal pad (stage 13-16). However, at moderate thresholds of AUROC (0.65), the number of significant expression domains from the heterogeneous network (47) is much more than from the homogenous "prot_domain" network (32). The "ChIP" only network was expected to outperform the "motif" only network because the ChIP data was from the corresponding developmental stage.

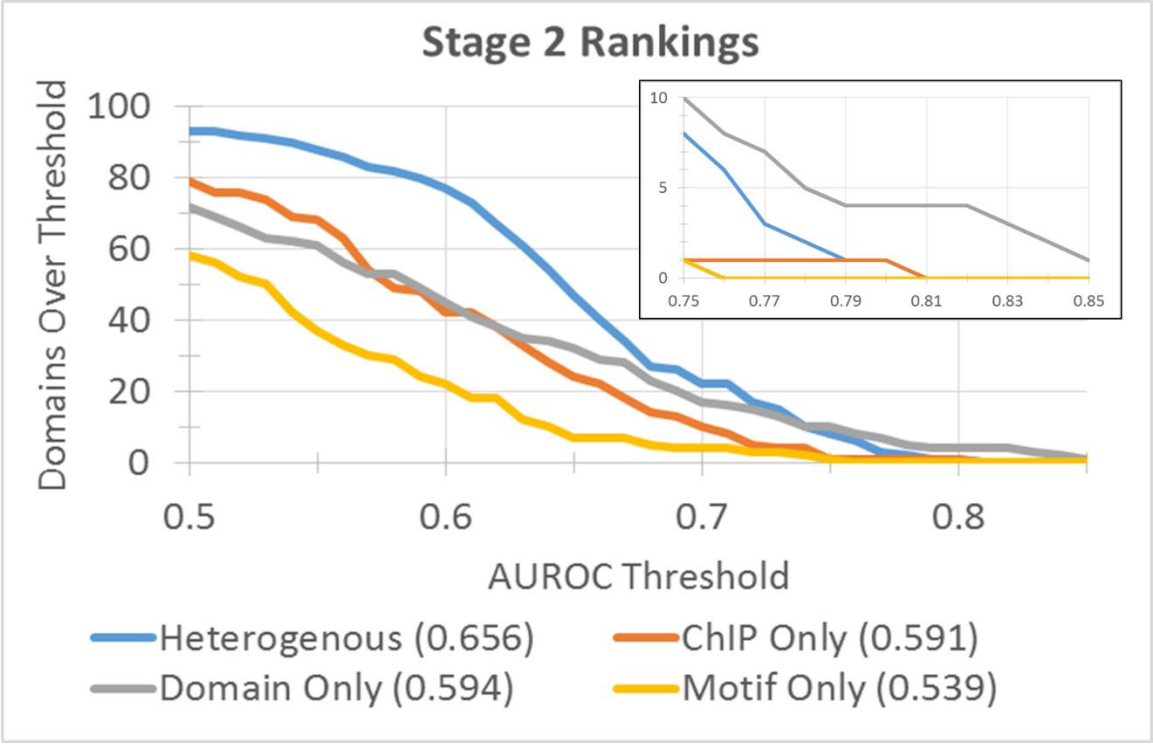


Figure 6.2 Comparison of RWR on Different Drosophila Networks. We compared the stage 2 rankings produced by our algorithm when the initial network was defined by single ("Domain", "ChIP", "Motif") or "Heterogeneous" feature types. We calculated the stage 2 AUROCs for each of the 92 expression domains and then plot the number of domains (y-axis) that were above each possible AUROC threshold (x-axis). The inset shows more detail for the chart region of high AUROC.

The first step of our procedure was to normalize the initial adjacency matrix by edge type to equalize the global probability that a walker follows a particular edge type. Without this normalization procedure, the average AUROC results of our two-stage method on the heterogeneous network are somewhat worse (0.646). We also examined the main parameter of the RWR method, the restart parameter, *c*. We ran the two-stage procedure on the heterogeneous network with six different values of the restart probability between 0 and 1. We found the best performance with the relatively high restart probability of 0.7 (Figure 6.3). The restart probability controls the influence of the network structure and the proximity of the query set on the final relevance vector. A high restart probability may be needed in the first stage to select relevant feature nodes that are more proximal to the query set than that are functioning as hubs in the network.

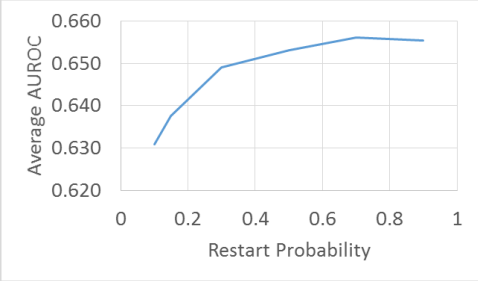


Figure 6.3 Effect of Restart Probability. For seven separate values of the restart probability (x-axis), we calculated the stage 2 AUROCs for each of the 92 expression domains and plotted the average value (y-axis).

6.4.2 Two stage RWR on multi-species networks

Our algorithm was designed to work with large, heterogeneous networks built from many public databases of biological knowledge. With improving high throughput sequencing techniques, the number of publicly available genomes is rapidly growing. We next sought to test whether including additional genomes in our biological network would improve ranking performance on the developmental gene sets. To this end, we constructed a "5 Insect" network with gene nodes representing genes from the fruit fly *D. melanogaster*, the mosquito *A. gambiae*, the honeybee *A. mellifera*, the jewel wasp *N. vitripennis*, and the beetle *T. castaneum*. As described in Section 6.2, the gene nodes within and between the five species were connected with weighted "homology" edges when they share high protein sequence according to BLAST. Additionally, all "prot_domain" and "motif" feature nodes were connected to gene nodes in all five species in the manner described in 6.2. Since the ChIP experiments were only available for *Drosophila*, the "ChIP" feature nodes only connect to fruit fly gene nodes. The new network had five times the number of species, but thirteen times the number of edges (Table 6.2). This was mostly due to the homology edges, which account for 78% of the edges in the "5 Insect" network.

		nEdges					
FeatureType	**nNodes**	**mosquito**	**fly**	**bee**	**wasp**	**beetle**	***Total***
homology	58,147	1,157,075	1,115,104	958,832	1,428,339	1,289,844	*5,949,194*
motif	222	285,774	165,577	208,123	339,611	429,526	*1,428,611*
prot_domain	3,671	31,372	34,244	31,865	33,701	35,039	*166,221*
ChIP	75		53,710				*53,710*
Grand Total	**62,115**	**1,474,221**	**1,368,635**	**1,198,820**	**1,801,651**	**1,754,409**	***7,597,736***

Table 6.2 Composition of 5 Insect Network. Lists the total number of nodes ("nNodes") of each feature type as well as the number of incoming edges ("nEdges") of each feature type for all genes nodes of a given insect in the multi-species, combined heterogeneous network.

Although there were 58,147 gene nodes, spanning five species, in this new network, our task was still to rank the 13,604 gene nodes in *Drosophila* for their relatedness to a specific developmental gene set. For this reason, the universe *U* of genes needed to calculate the "baseline ranking" in the first stage comprised only fruit fly genes. In this way, the baseline ranking shows the relevance of the features nodes with respect to the network and all fruit fly genes. This careful construction of the baseline ranking prevents features like the "ChIP" nodes that are *Drosophila* specific from always being selected as relevant features for the second stage simply because they are only connected to genes from the same species as the query genes. Apart from this modification, the two-stage RWR ranking algorithm and its evaluations were run on the "5 Insect" network in the same manner as *Drosophila* network discussed above. Because of the increased size of the data, number of iterations required to converge, and computational demands to perform the algorithm on the "5 Insect" network, we focused on only 12 of the 92 expression domains (Table 6.3).

Expression Domain	Stage	#Genes
brain primordium	11-12	779
dorsal ectoderm anlage in statu nascendi	4-6	255
dorsal ectoderm primordium	9-10	217
dorsal epidermis primordium	11-12	354
embryonic dorsal epidermis	13-16	714
embryonic ventral epidermis	13-16	638
procephalic ectoderm anlage in statu nascendi	4-6	229
procephalic ectoderm primordium	9-10	426
ventral ectoderm anlage in statu nascendi	4-6	221
ventral ectoderm primordium	9-10	287
ventral epidermis primordium	11-12	295
ventral nerve cord primordium	11-12	720

Table 6.3 List of Selected Expression Domains. Twelve expression domains from various developmental stages and gene set sizes were selected for additional analysis.

The average AUROC value for the stage 2 query rankings using the "5 Insect" heterogeneous network was higher (0.752) than the corresponding value on the *Drosophila* only heterogeneous network (0.728) (Figure 6.4). As before, the stage 2 rankings in the "5 Insect" heterogeneous network were also better than the stage 1 rankings. The improvement upon incorporating additional species was in addition to the improvement we observed with heterogeneous over homogenous networks. The "5 Insect" network contained many additional nodes and edges that do not directly relate to the fruit fly genes we are ranking. However, the

advantage of the network approach is that many indirect connections contribute meaningfully to the rankings. Presumably, more meaningful "motif" or "prot_domain" features will be conserved in multiple species and form dense subnetworks within our "5 Insect" heterogeneous network. The relevance of these nodes within dense subnetworks containing query genes will be additionally increased by the RWR algorithm, and the feature nodes corresponding to these conserved "motif" or "prot_domain" features will be ranked higher for selection for the second stage.

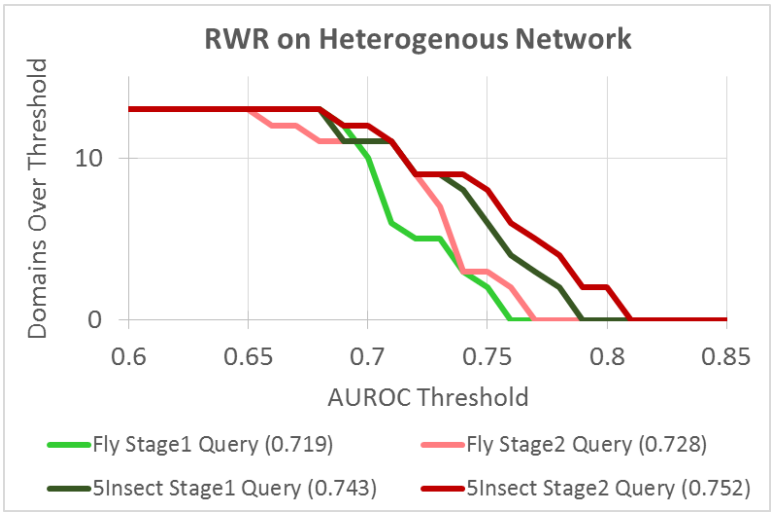


Figure 6.4 Comparison between Single and Multi-Species Networks. We compared the stage 1 and stage 2 rankings when the initial network was defined as the heterogeneous network either from a single species ("Fly") or from multiple species ("5Insect"). We calculated the AUROCs from each stage's rankings for each of the 12 selected expression domains and then plot the number of domains (y-axis) that were above each possible AUROC threshold (x-axis).

6.4.3 Query specific feature nodes

To create the query specific subnetwork for the second stage RWR, we identified the set of feature nodes that are the most specifically relevant to the query gene set. If there are k feature types, we select $50k$ feature nodes to be included in the subnetwork. Of the 150 features nodes selected from our heterogeneous *Drosophila* network, on average, 6 were "motif" nodes, 107 were "prot_domain" nodes, and 38 were "ChIP" nodes. This was a strong enrichment for ChIP feature nodes, which only account for 2% of all feature nodes. This enrichment is not surprising given that the ChIP features were derived from experiments performed in the same developmental stages as query gene sets. This was a crude confirmation that our feature selection procedure is selecting query relevant features. Some ChIP feature nodes were selected for many (>65) of the 92 different query gene sets. These nodes corresponded to the DNA-binding of

pioneer factors TRL and VFL or important developmental regulators, such as TWI, HB, and EVE. The zinc-finger, homeobox, and helix-loop-helix protein domains also appeared as selected features for more than 50 of the 92 expression domains. These were the most common DNA-binding protein domains, and their appearance on the list of most relevant features is consistent with the common knowledge that transcription factors are a key component of gene expression control during development. We also compared the feature selection results between "5 Insect" and "*Drosophila* only" evaluations. For the 12 selected expression domain gene sets, a total of 1800 features were selected in each evaluation. 1540 (85%) of the features selected in the "Fly" only analysis were also selected in the corresponding test using the multi-species network.

6.4.4 Comparison to GeneMANIA

We attempted to compare the performance of our two-stage random walk-based ranking procedure to the popular tool GeneMANIA. This tool implements label propagation on a gene-"gene affinity network" to rank genes on their similarity to a given set. The only data type in our previous analysis that has already been preprocessed into a GeneMANIA affinity network was the "prot_domain" feature type of Pfam domain annotations. In the GeneMANIA affinity network, two genes are joined if they share Pfam domains, but the number and types of the domains shared are lost in the collapsed one edge representation. In our network, we explicitly connected gene nodes that share protein domains to the same feature node (representing that protein domain) preserving the specific details of gene-gene relationships. Using 10% of each expression domains as test sets and the AUROC evaluation metric, we compared the GeneMANIA algorithm with its Pfam protein domain affinity network to our two-stage RWR method with the homogeneous "prot_domain" network in *Drosophila*. We found that our two-stage algorithm outperforms GeneMANIA at high values of AUROC threshold (Figure 6.5). For example, at an AUROC threshold of 0.7, significant rankings were discovered for 17 expression domains with our RWR procedure and for only 8 expression domains with GeneMANIA. Our algorithm was also able to return the most relevant protein domains, a capability that GeneMANIA lacks.

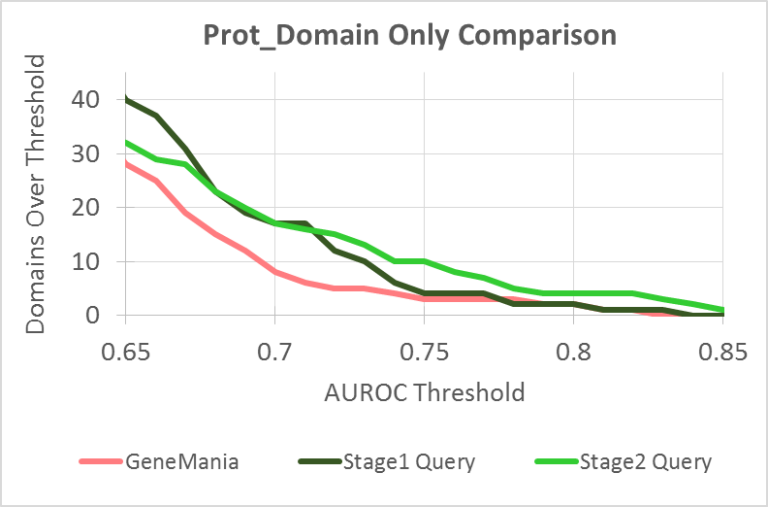


Figure 6.5 Comparison to GeneMANIA. We calculated the AUROCs produced from stage 1 and stage 2 rankings for the 92 expression domains when the initial network was constructed from only "Prot_Domain" feature nodes and *Drosophila* gene nodes. We also calculated comparable AUROCs for the same expression domains using the GeneMANIA algorithm and their affinity network defined from Pfam protein domain annotations. For each evaluation, we plot the number of domains (y-axis) that were above each possible AUROC threshold (x-axis).

6.5 Evaluations with multi-species behavioral aggression sets

Finally, we applied our algorithm to experimentally derived gene sets that are challenging to analyze with common existing tools. In a recent study [7], investigators attempted to understand if there are conserved neuromolecular mechanisms that underlie the common behavior of aggressive response to territorial intrusion in social animals. This study examined the transcriptomic state of brains in three greatly diverged social animals, the mouse *M. musculus*, the stickleback fish *G. aculeatus*, and the honeybee *A. mellifera*. The analysis in this paper, following the common analysis paradigm, separately examines differentially expressed (DE) genes in each species to find statistically significant Gene Ontology terms and *cis*-regulatory elements that are shared across species. Our method offers the potential for studying the DE gene sets from three species in an integrated framework that may enable more subtle signals of potential conserved genetic "toolkits" to reveal themselves.

6.5.1 Construction of aggression network and query sets

To construct the network for the analysis of this dataset, we incorporated heterogeneous information from all three of the species in the study as well as two additional, well-annotated species *D. melanogaster* and *H. sapiens*. We constructed a weighted network with nodes and edges described in detail in Section 6.2. We connected the gene nodes within and between species with "homology" edges defined from all-pairs BLAST results. We connected 3,671

"prot_domain" feature nodes to gene nodes in all five species based on the corresponding HMMER scans results. In this "aggression" network, we included "Gene Ontology" feature nodes for 1,827 GO terms with frequent annotation. For the human, mouse, and fly genes, we added "Gene Ontology" edges of weight 2 for curated GO annotations and weight 1 for inferred annotations. We do not include any edges between "Gene Ontology" feature nodes and genes nodes of the fish or bee because most of their GO annotations included in Ensembl [179] are derived from orthology. Finally, we add "brain atlas" nodes and edges that connected these feature nodes to mouse gene nodes that are specifically expressed in one of twelve defined brain regions. This new "aggression" network Table 6.4 has the same number of species as the "5 insect" network, but is 74% larger because of the greater number of vertebrate genes. This network is dominated by the homology edges, which account for 95% of all edges. Overall, there are 76,060 genes in the multi-species, heterogeneous aggression network and over 13 million edges.

		nEdges					
FeatureType	nNodes	human	fish	mouse	fly	bee	*Total*
homology	76,060	2,336,732	3,876,000	3,036,361	1,712,592	1,609,365	*12,571,050*
prot_domain	3,671	71,324	64,373	71,795	34,244	31,865	*273,601*
Gene Ontology	1,827	182,858		171,027	51,515		*405,400*
Brain Atlas	12			1,086			*1,086*
Grand Total	**81,570**	**2,590,914**	**3,940,373**	**3,280,269**	**1,798,351**	**1,641,230**	***13,251,137***

Table 6.4 Composition of Aggression Network. Lists the total number of nodes ("nNodes") of each feature type as well as the number of incoming edges ("nEdges") of each feature type for all genes nodes of a given species in the multi-species, combined heterogeneous network.

We obtained one gene set of differentially expressed (DE) genes from each species from the aggression study [7]. At a FDR of 0.1, they report 153 bee genes, 499 fish genes, and 883 mouse genes to be differentially expressed in the brains of the social animals when exposed to an intruder. In this analysis, we were interested in ranking genes and features for their relatedness to all three DE gene sets simultaneously. To this end, we created a "3 species" gene set from the combination of all 1,535 DE genes. For each of these four DE gene sets, we created an appropriate gene universe set (genes that need to be ranked by our procedure). The gene universe set that corresponded to each DE gene set was defined as all of the genes from only the corresponding species. This means that although there are five species represented in the "aggression" network, we were not interested in ranking the genes of the fruit fly or human.

6.5.2 Aggression related features

We ran our two-stage RWR pipeline with the "3 species" DE query gene set and the heterogeneous, multi-species aggression network. We found that many of the feature nodes that most specifically related to the query set of DE genes (greatest value of $\tilde{\mathbf{v}}_{Q} - \tilde{\mathbf{v}}_{U}$) and report the top ten in Table 6.5.

Rank	Feature Node	Feature Type
1	Striatum	Brain Atlas
2	Retrohippocampal	Brain Atlas
3	Hippocampus	Brain Atlas
4	Pallidum	Brain Atlas
5	MRJP	Prot_domain
6	PMP22_Claudin	Prot_domain
7	JHBP	Prot_domain
8	Olfactory	Brain Atlas
9	Globin	Prot_domain
10	Claudin_2	Prot_domain

Table 6.5 Ten Query Specific Features The top ten feature nodes selected with our algorithm on the "3 species" query set and multi-species, heterogeneous aggression network. Each node is listed along with its feature type.

In particular, the feature node corresponding to the "Striatum" brain region was ranked first. This is consistent with the striatum being the part of the brain responsible for coordinating movement with motivation, an important component of an aggressive behavior response to an intruder. It has been demonstrated that damage to the striatum can result is aberrant social behavior [181]. The next most relevant feature nodes include the retrohippocampus, the hippocampus, and the pallidum, which are known to be involved in emotions and movement or motivation and behavior. We also found the protein domain feature nodes for major royal jelly protein (MRJP), juvenile hormone binding protein (JHBP) in our top ten list. Genes containing the MRJP domain have been previously implicated in behavior because of their expression in the mushroom bodies of honeybee brains [182, 183]. JHBP domain genes have also been correlated with hygienic behaviors in honeybees in response to infestations of parasitic mites [184]. There were several "Gene Ontology" features identified by our method as relevant to our "3 species" DE query set that were ranked in the top forty feature nodes. These included terms involving the plasma membrane, protein binding, and ribosome. The fifth most related Gene Ontology feature node was for the term "Hormone activity", which was also discovered in the original study [7].

6.5.3 Observations about gene rankings in aggression study

Table 6.5 shows AUROC evaluations on the 10% held out test sets on the stage two rankings produced by our algorithm in the aggression study. For ranking the DE query set defined from three species ("3 species" column of Table 6.6), we found that the heterogeneous multi-species network (AUROC 6.97) perform better than any homogeneous, multispecies network containing a single feature type. Our method successfully enabled us to integrate experimental results from different species with knowledge from many different sources in a single framework.

		DE Gene Sets			
Network	Features	3 Species	Mouse	Bee	Fish
Multi-species	Heterogeneous	0.690	0.788	0.595	0.647
	Prot Domain	0.567	0.611	0.692	0.627
	Brain Atlas	0.556	0.556		
	Gene Ontology	0.568	0.568		
Single Species	Heterogeneous		0.631	0.696	0.651

Table 6.6 AUROCs in Aggression Network. The query sets are differently expressed genes in a single species or the combination across all three studied species. The initial network is defined either from the single species that matches the query set or from all five species. It is also the heterogeneous combination of all feature types or a network containing only edges of a single feature type. Combinations of initial networks and query sets that were not examined are reported as gray cells.

We also examined the DE gene set of each species separately to check if the DE gene sets have varying levels of coherence that may make it more or less difficult to identify related genes. For each species, we tested their specific DE genes within our multi-species network as well as networks constructed by extracting all edges that connect to gene nodes of that single species. In general, we found that the DE gene sets of species that perform poorly in their single species networks show the greatest improvement when using the multi-species networks. In particular, we poorly ranked the mouse DE genes in the mouse single species heterogeneous network. However, when incorporating information from additional species, we see a great improvement (AUROC in heterogeneous, multi-species network 0.762). We also tested whether including the computationally inferred "Gene Ontology" edges improved or worsened the ranking predictions. In all cases, inclusion of the inferred edges enabled better rankings, suggesting future techniques in constructing these heterogeneous networks may choose to instantiate specific inferred relationships to poorly annotated genomes from orthology rather than relying on the information to propagate through homology edges.

6.6 Discussion

We have developed a method to rank genes for their relatedness to a given set in the context of large, heterogeneous information represented as a network. We have shown that the rankings improve when more sources of information are incorporated into the network and even when data from additional species are appended. Our algorithm applies a two-stage RWR to rank related genes and, as a byproduct, produces a list of features that are specifically related to the gene set. We have shown its application in examining embryonic expression domains in *Drosophila* and transcriptomic responses to intruders in a cross-species study.

One of the driving reasons for selecting a random walk with restart approach is scalability. With genome sequencing projects like the 10,000 Vertebrate Genomes (Genomes 10k) and 5000 Insect Genomes (i5k) underway and high throughput technologies becoming less expensive and more efficient, a biological network containing all public data would need to scale to thousands of species, covering tens of millions of genes and potentially billions of functional interactions. One common approach to address computational scalability is the paradigm of data and computation distribution offered by MapReduce [185]. The reliability and efficiency of this framework has led to its widespread adoption, and public instances (e.g. the Amazon Elastic Compute Cloud) provide a platform for users to store large networks and deploy analysis tools on them. We chose to implement a message passing based Random Walk with Restart (RWR) algorithm for our functional annotation tool because this algorithm easily maps to a MapReduce framework. The RWR algorithm implemented in the graph mining software PEGASUS [186] has been shown to scale to graphs with billions of nodes and edges. More recent software, B_LIN [180], Pregel [187], GraphLab [188], and GraphX [189], are explicitly designed to improve performance in scalable graph processing by carefully distributing data and minimizing communication costs. In the 80K node, 13M edge multi-species heterogeneous aggression network runs presented in this chapter, representing the data required at least 4 GB of RAM and processing it took several hours. At these requirements, it becomes difficult to optimize the restart parameter or the number of selected features in the second stage subnetwork for each query set. Scalability is of utmost importance since all of our results suggest that the algorithm is able to produce the best rankings when given the largest, most diverse initial network.

There are several limitations to the random walk based approach. First, we are only able to represent positive information. Edges are only able to convey how closely related two nodes

are and nodes are only allowed to be annotated as belonging to the given gene set. However, negative information may perhaps create a more nuanced network and produce better outcomes. For example, we may want to add edges that represent mutual exclusivity or strong anti-correlation between two nodes in the network. We may also have negative examples of our gene set property of interest that we would like to annotate and incorporate to make rankings more accurate. Many of these properties may be addressed by remapping our random walk on a connectivity network algorithm into an application of belief propagation on probabilistic graphical models [190, 191]. Additionally, although we normalize our edges by type, the RWR does specifically treat different types of edges in a distinguishable way. Some studies have attempted to control how information is passed through different edge types by defining specific meta-paths [192] that dictate a sequence of node types that must be followed to inform a relationship between two nodes. Our simple, two-stage RWR algorithm for gene ranking provides a solution to and highlights the challenges of performing analysis of experimental data on massive, heterogeneous networks of biological knowledge.

7 CONCLUSION

Analysis of co-expressed gene sets is fundamental to genomic research. Especially important are analyses that attempt to understand the transcription factors and regulatory enhancer sequences that are crucial in the set of genes sharing a similar expression pattern. These examinations into transcriptional regulation provide the investigator with insights into the signals and mechanisms that affect the outcomes in their cell type of interest. In this dissertation, we have attempted to 1) provide this type of regulatory analysis using TF motif based methods with limited additional experimental data and 2) demonstrate network based methods that integrate regulatory analysis with other heterogeneous data types in a single framework. The specific contributions of the dissertation were:

1. We developed a pipeline for identifying putative regulators of a co-expressed gene set. Because our method relies on TF motifs, we are able to investigate the regulatory potential of hundreds of transcription factors. This is in contrast to alternative ChIP based methods that are only possible for a limited number of TFs in a limited number of well-studied tissues. We demonstrated the effectiveness of incorporating a novel sequence conservation score and data from a single experimental assay of chromatin accessibility into our pipeline. After applying our methods to hundreds of expression domain in *Drosophila* embryos, we discovered novel insights into developmental regulators and tissues. We also demonstrated the applicability of the pipeline to several other species and cell types and made available a real-time web tool for this type of analysis.
2. We also developed a method for identifying these regulatory signals when they are shared across genes from multiple experimental conditions, tissues, or species. Our *cis*-Metalysis program incorporates a novel statistical procedure for combining independent p-values from multiple tests with only an unknown subset expected bear evidence of a signal. It also is one of few frameworks that are able to systematically test for regulatory signatures of combinations of TFs, a common feature of transcriptional regulation of gene expression in eukaryotes. We applied the method to gene sets relating to honeybee behavioral maturation and were able to discover informative results after applying strict EVD-based significance criteria.

3. We demonstrated that models of enhancer activity are effective in annotating enhancers of genes expressed in a cell type. For the complex patterns in the A/P segmentation system, we developed a "pattern generating potential" similarity measure that identifies when putative enhancers capture discrete aspects of their target gene's expression. With PGP and our enhancer model, we were also able to identify a network of regulatory edges that were well supported by literature and more consistent with known regulatory roles of TFs than one produced from ChIP data. We applied our enhancer modeling and annotation strategy to 195 poorly characterized expression domains of fly embryonic development by creating "noisy" sets of training enhancers and identifying putative regulators of each cell type. With motif scores combined with a single chromatin accessibility experimentally assay, we were able to identify enhancers in more expression domains than any other method.
4. Finally, we designed a novel algorithm that, given a heterogeneous network containing a large amount of biological knowledge and a co-expressed query gene set, identifies the most related genes and biological properties. Our tests show that the results of this algorithm improve as more information and additional species are added to the original network. We also demonstrated its applicability to an aggression study involving three separate species where standard methods do not apply.

In the future, we expect high throughput characterization assays of the non-coding genome to improve and become more cost effective. However, until there is a more cost-effective method for characterizing TF binding than ChIP assays, we expect that there will remain great value in TF motif based methods for their approximation. Driven by the rise of high throughput technologies, gene set analysis in the context of "big data" heterogeneous, biological networks is an emerging topic. Principles from social networks and recommendation systems must be integrated into biological settings to guide researchers and physicians in their investigations into drivers of disease and in their treatment of patients. Additional network-aided analysis tasks beyond gene ranking (such as classification and clustering) will enable even further understanding of related transcriptomic profiles.

REFERENCES

1. Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM: **Global analysis of patterns of gene expression during *Drosophila* embryogenesis.** *Genome Biol* 2007, **8:**R145.
2. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
3. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57-63.
4. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, et al: **Genome-wide atlas of gene expression in the adult mouse brain.** *Nature* 2007, **445:**168-176.
5. Zou Z, Saha TT, Roy S, Shin SW, Backman TW, Girke T, White KP, Raikhel AS: **Juvenile hormone and its receptor, methoprene-tolerant, control the dynamics of mosquito gene expression.** *Proc Natl Acad Sci U S A* 2013, **110:**E2173-2181.
6. Ament SA, Wang Y, Chen CC, Blatti CA, Hong F, Liang ZS, Negre N, White KP, Rodriguez-Zas SL, Mizzen CA, et al: **The transcription factor ultraspiracle influences honey bee social behavior and behavior-related gene expression.** *PLoS Genet* 2012, **8:**e1002596.
7. Rittschof CC, Bukhari SA, Sloofman LG, Troy JM, Caetano-Anolles D, Cash-Ahmed A, Kent M, Lu X, Sanogo YO, Weisner PA, et al: **Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee.** *Proc Natl Acad Sci U S A* 2014.
8. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nat Genet* 2004, **36:**1090-1098.
9. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37:**1-13.
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
11. Tanabe M, Kanehisa M: **Using the KEGG database resource.** *Curr Protoc Bioinformatics* 2012, **Chapter 1:**Unit1 12.
12. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.
13. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.
14. Yanez-Cuna JO, Kvon EZ, Stark A: **Deciphering the transcriptional *cis*-regulatory code.** *Trends Genet* 2013, **29:**11-22.
15. Zambelli F, Prazzoli GM, Pesole G, Pavesi G: **Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets.** *Nucleic Acids Res* 2012, **40:**W510-515.

16. Auerbach RK, Chen B, Butte AJ: **Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool.** *Bioinformatics* 2013, **29:**1922-1924.
17. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al: **A *cis*-regulatory map of the *Drosophila* genome.** *Nature* 2011, **471:**527-531.
18. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching.** *Nucleic Acids Res* 2009, **37:**W202-208.
19. Hertz GZ, Hartzell GW, 3rd, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Comput Appl Biosci* 1990, **6:**81-92.
20. Simcha D, Price ND, Geman D: **The limits of de novo DNA motif discovery.** *PLoS One* 2012, **7:**e47836.
21. Wenger AM, Clarke SL, Guturu H, Chen J, Schaar BT, McLean CY, Bejerano G: **PRISM offers a comprehensive genomic approach to transcription factor function prediction.** *Genome Res* 2013, **23:**889-904.
22. Herrmann C, Van de Sande B, Potier D, Aerts S: **i-cisTarget: an integrative genomics method for the prediction of regulatory features and *cis*-regulatory modules.** *Nucleic Acids Res* 2012, **40:**e114.
23. Li XY, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD: **The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding.** *Genome Biol* 2011, **12:**R34.
24. Whitney O, Pfenning AR, Howard JT, Blatti CA, Liu F, Ward JM, Wang R, Audet J-N, Kellis M, Mukherjee S, et al: **Core and region-enriched networks of behaviorally regulated genes and the singing genome.** *Science* 2014, **346**.
25. Sanogo YO, Band M, Blatti C, Sinha S, Bell AM: **Transcriptional regulation of brain gene expression in response to a territorial intrusion.** *Proc Biol Sci* 2012, **279:**4929-4938.
26. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data.** *Bioinformatics* 2006, **22:**1122-1129.
27. Tanay A, Sharan R, Kupiec M, Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data.** *Proc Natl Acad Sci U S A* 2004, **101:**2981-2986.
28. Halperin Y, Linhart C, Ulitsky I, Shamir R: **Allegro: analyzing expression and sequence in concert to discover regulatory programs.** *Nucleic Acids Res* 2009, **37:**1566-1579.
29. Huttenhower C, Mutungu KT, Indik N, Yang W, Schroeder M, Forman JJ, Troyanskaya OG, Coller HA: **Detailing regulatory networks through large scale data integration.** *Bioinformatics* 2009, **25:**3267-3274.
30. Whitlock MC: **Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach.** *J Evol Biol* 2005, **18:**1368-1373.
31. Chang LC, Lin HM, Sibille E, Tseng GC: **Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline.** *BMC Bioinformatics* 2013, **14:**368.

32. Shlyueva D, Stampfel G, Stark A: **Transcriptional enhancers: from properties to genome-wide predictions.** *Nat Rev Genet* 2014, **15:**272-286.
33. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al: **An atlas of active enhancers across human cell types and tissues.** *Nature* 2014, **507:**455-461.
34. Frith MC, Li MC, Weng Z: **Cluster-Buster: Finding dense clusters of motifs in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3666-3668.
35. Sinha S, Liang Y, Siggia E: **Stubb: a program for discovery and analysis of *cis*-regulatory modules.** *Nucleic Acids Res* 2006, **34:**W555-559.
36. Kazemian M, Zhu Q, Halfon MS, Sinha S: **Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison.** *Nucleic Acids Res* 2011, **39:**9463-9472.
37. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization.** *Nat Methods* 2012, **9:**215-216.
38. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE: **Combinatorial binding predicts spatio-temporal *cis*-regulatory activity.** *Nature* 2009, **462:**65-70.
39. Wilczynski B, Liu YH, Yeo ZX, Furlong EE: **Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state.** *PLoS Comput Biol* 2012, **8:**e1002798.
40. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al: **Identification of functional elements and regulatory circuits by *Drosophila* modENCODE.** *Science* 2010, **330:**1787-1797.
41. Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M: **Predictive regulatory models in *Drosophila* melanogaster by integrative inference of transcriptional networks.** *Genome Res* 2012, **22:**1334-1349.
42. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al: **ChIP-seq accurately predicts tissue-specific activity of enhancers.** *Nature* 2009, **457:**854-858.
43. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.** *Genome Res* 2011, **21:**447-455.
44. He X, Samee MA, Blatti C, Sinha S: **Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression.** *PLoS Comput Biol* 2010, **6:**epublish.
45. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U: **Predicting expression patterns from regulatory sequence in *Drosophila* segmentation.** *Nature* 2008, **451:**535-540.
46. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117:**185-198.
47. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I: **Predicting tissue-specific enhancers in the human genome.** *Genome Res* 2007, **17:**201-211.
48. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U: **Predicting cell-type-specific gene expression from regions of open chromatin.** *Genome Res* 2012, **22:**1711-1722.
49. Chen X, Blanchette M: **Prediction of tissue-specific *cis*-regulatory modules using Bayesian networks and regression trees.** *BMC Bioinformatics* 2007, **8 Suppl 10:**S2.

50. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U: **Transcriptional control in the segmentation gene network of Drosophila.** *PLoS Biol* 2004, **2:**E271.
51. Cornish AJ, Markowetz F: **SANTA: quantifying the functional content of molecular networks.** *PLoS Comput Biol* 2014, **10:**e1003808.
52. Pena-Castillo L, Tasan M, Myers CL, Lee H, Joshi T, Zhang C, Guan Y, Leone M, Pagnani A, Kim WK, et al: **A critical assessment of Mus musculus gene function prediction using integrated genomic evidence.** *Genome Biol* 2008, **9 Suppl 1:**S2.
53. Hou JP, Ma J: **DawnRank: discovering personalized driver genes in cancer.** *Genome Med* 2014, **6:**56.
54. Hofree M, Shen JP, Carter H, Gross A, Ideker T: **Network-based stratification of tumor mutations.** *Nat Methods* 2013, **10:**1108-1115.
55. Chen X, Liu MX, Yan GY: **Drug-target interaction prediction by random walk on the heterogeneous network.** *Mol Biosyst* 2012, **8:**1970-1978.
56. Li Y, Patra JC: **Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network.** *Bioinformatics* 2010, **26:**1219-1224.
57. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.** *Nucleic Acids Res* 2010, **38:**W214-220.
58. Davidson E: *Genomic Regulatory Systems: Development and Evolution.* Academic Press; 2001.
59. Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, Halfon MS: **REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila.** *Nucleic Acids Res* 2011, **39:**D118-123.
60. Visel A, Minovitsky S, Dubchak I, Pennacchio LA: **VISTA Enhancer Browser--a database of tissue-specific human enhancers.** *Nucleic Acids Res* 2007, **35:**D88-92.
61. Arnone MI, Davidson EH: **The hardwiring of development: organization and function of genomic regulatory systems.** *Development* 1997, **124:**1851-1864.
62. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489:**75-82.
63. Zaret KS, Carroll JS: **Pioneer transcription factors: establishing competence for gene expression.** *Genes Dev* 2011, **25:**2227-2241.
64. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A: **Genome-wide quantitative enhancer activity maps identified by STARR-seq.** *Science* 2013, **339:**1074-1077.
65. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome.** *Cell* 2008, **132:**311-322.
66. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al: **An expansive human regulatory lexicon encoded in transcription factor footprints.** *Nature* 2012, **489:**83-90.
67. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD: **FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin.** *Genome Res* 2007, **17:**877-885.

68. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316:**1497-1502.
69. Rhee HS, Pugh BF: **Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution.** *Cell* 2011, **147:**1408-1419.
70. Kvon EZ, Stampfel G, Yanez-Cuna JO, Dickson BJ, Stark A: **HOT regions function as patterned developmental enhancers and have a distinct *cis*-regulatory signature.** *Genes Dev* 2012, **26:**908-913.
71. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, et al: **Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm.** *PLoS Biol* 2008, **6:**e27.
72. Ernst J, Kellis M: **Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types.** *Genome Res* 2013, **23:**1142-1154.
73. Newburger DE, Bulyk ML: **UniPROBE: an online database of protein binding microarray data on protein-DNA interactions.** *Nucleic Acids Res* 2009, **37:**D77-82.
74. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al: **DNA-binding specificities of human transcription factors.** *Cell* 2013, **152:**327-339.
75. Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA: **A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system.** *Nucleic Acids Res* 2008, **36:**2547-2560.
76. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al: **Determination and inference of eukaryotic transcription factor sequence specificity.** *Cell* 2014, **158:**1431-1443.
77. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D, et al: **Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium.** *Nucleic Acids Res* 2013, **41:**D171-176.
78. Kim J, Cunningham R, James B, Wyder S, Gibson JD, Niehuis O, Zdobnov EM, Robertson HM, Robinson GE, Werren JH, Sinha S: **Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances.** *PLoS Comput Biol* 2010, **6:**e1000652.
79. Morozov AV, Siggia ED: **Connecting protein structure with predictions of regulatory sites.** *Proc Natl Acad Sci U S A* 2007, **104:**7068-7073.
80. Negre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, Feng X, Ahmad K, Russell S, White RA, et al: **A comprehensive map of insulator elements for the *Drosophila* genome.** *PLoS Genet* 2010, **6:**e1000814.
81. van Steensel B, Dekker J: **Genomics tools for unraveling chromosome architecture.** *Nat Biotechnol* 2010, **28:**1089-1095.
82. Brody T: **The Interactive Fly: gene networks, development and the Internet.** *Trends Genet* 1999, **15:**333-334.
83. Blatti C, Sinha S: **Motif enrichment tool.** *Nucleic Acids Res* 2014, **42:**W20-25.
84. Kazemian M, Blatti C, Richards A, McCutchan M, Wakabayashi-Ito N, Hammonds AS, Celniker SE, Kumar S, Wolfe SA, Brodsky MH, Sinha S: **Quantitative analysis of the *Drosophila* segmentation regulatory network using pattern generating potentials.** *PLoS Biol* 2010, **8:**epublish.

85. Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J: **RSAT 2011: regulatory sequence analysis tools.** *Nucleic Acids Res* 2011, **39:**W86-91.
86. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, et al: **ENCODE data in the UCSC Genome Browser: year 5 update.** *Nucleic Acids Res* 2013, **41:**D56-63.
87. McLeay RC, Bailey TL: **Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data.** *BMC Bioinformatics* 2010, **11:**165.
88. Whitfield CW, Ben-Shahar Y, Brillet C, Leoncini I, Crauser D, Leconte Y, Rodriguez-Zas S, Robinson GE: **Genomic dissection of behavioral maturation in the honey bee.** *Proc Natl Acad Sci U S A* 2006, **103:**16068-16075.
89. Cheng Q, Kazemian M, Pham H, Blatti C, Celniker SE, Wolfe SA, Brodsky MH, Sinha S: **Computational identification of diverse mechanisms underlying transcription factor-DNA occupancy.** *PLoS Genet* 2013, **9:**e1003571.
90. Kaplan T, Li XY, Sabo PJ, Thomas S, Stamatoyannopoulos JA, Biggin MD, Eisen MB: **Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development.** *PLoS Genet* 2011, **7:**e1001290.
91. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.** *Nat Methods* 2009, **6:**283-289.
92. Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL: **Epigenetic priors for identifying active transcription factor binding sites.** *Bioinformatics* 2012, **28:**56-62.
93. Whitington T, Perkins AC, Bailey TL: **High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites.** *Nucleic Acids Res* 2009, **37:**14-25.
94. He X, Chen CC, Hong F, Fang F, Sinha S, Ng HH, Zhong S: **A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data.** *PLoS One* 2009, **4:**e8155.
95. Won KJ, Ren B, Wang W: **Genome-wide prediction of transcription factor binding sites using an integrated model.** *Genome Biol* 2010, **11:**R7.
96. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
97. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic Acids Res* 2010, **38:**D105-110.
98. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:**374-378.
99. Zhu LJ, Christensen RG, Kazemian M, Hull CJ, Enuameh MS, Basciotta MD, Brasefield JA, Zhu C, Asriyan Y, Lapointe DS, et al: **FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system.** *Nucleic Acids Res* 2011, **39:**D111-117.

100. Warner JB, Philippakis AA, Jaeger SA, He FS, Lin J, Bulyk ML: **Systematic identification of mammalian regulatory motifs' target genes and functions.** *Nat Methods* 2008, **5:**347-353.
101. Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19 Suppl 1:**i292-301.
102. Felsenstein J: **Maximum-likelihood estimation of evolutionary trees from continuous characters.** *Am J Hum Genet* 1973, **25:**471-492.
103. Stone EA, Sidow A: **Constructing a meaningful evolutionary average at the phylogenetic center of mass.** *BMC Bioinformatics* 2007, **8:**222.
104. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al: **Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures.** *Nature* 2007, **450:**219-232.
105. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.** *Nat Methods* 2013, **10:**1213-1218.
106. Thomas S, Li XY, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, et al: **Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development.** *Genome Biol* 2011, **12:**R43.
107. MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keranen SV, 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* 2009, **10:**R80.
108. Busser BW, Huang D, Rogacki KR, Lane EA, Shokri L, Ni T, Gamble CE, Gisselbrecht SS, Zhu J, Bulyk ML, et al: **Integrative analysis of the zinc finger transcription factor Lame duck in the *Drosophila* myogenic gene regulatory network.** *Proc Natl Acad Sci U S A* 2012, **109:**20768-20773.
109. Nien CY, Liang HL, Butcher S, Sun Y, Fu S, Gocha T, Kirov N, Manak JR, Rushlow C: **Temporal coordination of gene networks by Zelda in the early *Drosophila* embryo.** *PLoS Genet* 2011, **7:**e1002339.
110. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, van Lohuizen M, Tanay A, Cavalli G: **Functional anatomy of polycomb and trithorax chromatin landscapes in *Drosophila* embryos.** *PLoS Biol* 2009, **7:**e13.
111. Adkins NL, Hagerman TA, Georgel P: **GAGA protein: a multi-faceted transcription factor.** *Biochem Cell Biol* 2006, **84:**559-567.
112. Harrison MM, Li XY, Kaplan T, Botchan MR, Eisen MB: **Zelda binding in the early *Drosophila* melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition.** *PLoS Genet* 2011, **7:**e1002266.
113. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM: **Systematic determination of patterns of gene expression during *Drosophila* embryogenesis.** *Genome Biol* 2002, **3:**RESEARCH0088.
114. McQuilton P, St Pierre SE, Thurmond J: **FlyBase 101--the basics of navigating FlyBase.** *Nucleic Acids Res* 2012, **40:**D706-714.
115. Mulholland NM, King IF, Kingston RE: **Regulation of Polycomb group complexes by the sequence-specific DNA binding proteins Zeste and GAGA.** *Genes Dev* 2003, **17:**2741-2746.

116. Orsi GA, Kasinathan S, Hughes KT, Saminadin-Peter S, Henikoff S, Ahmad K: **High-resolution mapping defines the cooperative architecture of Polycomb response elements.** *Genome Res* 2014, **24:**809-820.
117. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, et al: **Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.** *Genome Res* 2006, **16:**656-668.
118. Kostyuchenko M, Savitskaya E, Koryagina E, Melnikova L, Karakozova M, Georgiev P: **Zeste can facilitate long-range enhancer-promoter communication and insulator bypass in *Drosophila* melanogaster.** *Chromosoma* 2009, **118:**665-674.
119. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, et al: **The genome of a songbird.** *Nature* 2010, **464:**757-762.
120. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif.** *Bioinformatics* 2011, **27:**1017-1018.
121. Ament SA, Blatti CA, Alaux C, Wheeler MM, Toth AL, Le Conte Y, Hunt GJ, Guzman-Novoa E, Degrandi-Hoffman G, Uribe-Rubio JL, et al: **New meta-analysis tools reveal common transcriptional regulatory basis for multiple determinants of behavior.** *Proc Natl Acad Sci U S A* 2012, **109:**E1801-1810.
122. West-Eberhard MJ: *Developmental plasticity and evolution.* Oxford ; New York: Oxford University Press; 2003.
123. Dettmann CP, Georgiou O: **Product of n independent uniform random variables.** *Statistics & Probability Letters* 2009, **79:**2501-2503.
124. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100:**9440-9445.
125. Ng P, Keich U: **GIMSAN: a Gibbs motif finder with significance analysis.** *Bioinformatics* 2008, **24:**2256-2257.
126. Alaux C, Sinha S, Hasadsri L, Hunt GJ, Guzman-Novoa E, DeGrandi-Hoffman G, Uribe-Rubio JL, Southey BR, Rodriguez-Zas S, Robinson GE: **Honey bee aggression supports a link between gene regulation and behavioral evolution.** *Proc Natl Acad Sci U S A* 2009, **106:**15400-15405.
127. Chandrasekaran S, Ament SA, Eddy JA, Rodriguez-Zas SL, Schatz BR, Price ND, Robinson GE: **Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states.** *Proc Natl Acad Sci U S A* 2011, **108:**18020-18025.
128. Sinha S, Adler AS, Field Y, Chang HY, Segal E: **Systematic functional characterization of *cis*-regulatory motifs in human core promoters.** *Genome Res* 2008, **18:**477-488.
129. Pick L: **Segmentation: painting stripes from flies to vertebrates.** *Dev Genet* 1998, **23:**1-10.
130. Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczynski B, Riddell A, Furlong EE: **Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development.** *Nat Genet* 2012, **44:**148-156.

131. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B: **A high-resolution map of the three-dimensional chromatin interactome in human cells.** *Nature* 2013, **503:**290-294.
132. McKay DJ, Lieb JD: **A common set of DNA regulatory elements shapes *Drosophila* appendages.** *Dev Cell* 2013, **27:**306-318.
133. Chen J, Hu Z, Phatak M, Reichard J, Freudenberg JM, Sivaganesan S, Medvedovic M: **Genome-wide signatures of transcription factor activity: connecting transcription factors, disease, and small molecules.** *PLoS Comput Biol* 2013, **9:**e1003198.
134. Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA: **Integrating diverse datasets improves developmental enhancer prediction.** *PLoS Comput Biol* 2014, **10:**e1003677.
135. Kvon EZ, Kazmar T, Stampfel G, Yanez-Cuna JO, Pagani M, Schernhuber K, Dickson BJ, Stark A: **Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo.** *Nature* 2014.
136. Poustelnikova E, Pisarev A, Blagov M, Samsonova M, Reinitz J: **A database for management of gene expression data in situ.** *Bioinformatics* 2004, **20:**2212-2221.
137. R Development Core Team: **R: A Language and Environment for Statistical Computing.** Vienna, Austria: R Foundation for Statistical Computing; 2008.
138. Perkins TJ, Jaeger J, Reinitz J, Glass L: **Reverse engineering the gap gene network of *Drosophila* melanogaster.** *PLoS Comput Biol* 2006, **2:**e51.
139. Zhu W, Foehr M, Jaynes JB, Hanes SD: *Drosophila* **SAP18, a member of the Sin3/Rpd3 histone deacetylase complex, interacts with Bicoid and inhibits its activity.** *Dev Genes Evol* 2001, **211:**109-117.
140. Lohr U, Chung HR, Beller M, Jackle H: **Antagonistic action of Bicoid and the repressor Capicua determines the spatial limits of *Drosophila* head gene expression domains.** *Proc Natl Acad Sci U S A* 2009.
141. Ochoa-Espinosa A, Yu D, Tsirigos A, Struffi P, Small S: **Anterior-posterior positional information in the absence of a strong Bicoid gradient.** *Proc Natl Acad Sci U S A* 2009, **106:**3823-3828.
142. Ludwig MZ, Patel NH, Kreitman M: **Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change.** *Development* 1998, **125:**949-958.
143. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE: **Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila* melanogaster and *Drosophila* pseudoobscura.** *Genome Biol* 2004, **5:**R61.
144. Kumar S, Consortium F: **A Knowledgebase Spatiotemporal Expression Patterns at a Genomic-scale in the Fruit-fly Embryogenesis (www.flyexpress.net). Arizona State University, Tempe, Arizona 85287, USA.** 2009.
145. Hong JW, Hendrix DA, Levine MS: **Shadow enhancers as a source of evolutionary novelty.** *Science* 2008, **321:**1314.
146. Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED: **Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in Drosophila.** *BMC Bioinformatics* 2004, **5:**129.

147. Pfeiffer BD, Jenett A, Hammonds AS, Ngo TT, Misra S, Murphy C, Scully A, Carlson JW, Wan KH, Laverty TR, et al: **Tools for neuroanatomy and neurogenetics in Drosophila.** *Proc Natl Acad Sci U S A* 2008, **105:**9715-9720.
148. Bergman CM, Carlson JW, Celniker SE: *Drosophila* **DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila* melanogaster.** *Bioinformatics* 2005, **21:**1747-1749.
149. Zhou B, Bagri A, Beckendorf SK: **Salivary gland determination in Drosophila: a salivary-specific, fork head enhancer integrates spatial pattern and allows fork head autoregulation.** *Dev Biol* 2001, **237:**54-67.
150. Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, Myasnikova E, Surkova S, Vanario-Alonso CE, Samsonova M, Sharp DH, Reinitz J: **Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila* melanogaster.** *Genetics* 2004, **167:**1721-1737.
151. Manu, Surkova S, Spirov AV, Gursky VV, Janssens H, Kim AR, Radulescu O, Vanario-Alonso CE, Sharp DH, Samsonova M, Reinitz J: **Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation.** *PLoS Biol* 2009, **7:**e1000049.
152. Stevens LM, Beuchle D, Jurcsak J, Tong X, Stein D: **The *Drosophila* embryonic patterning determinant torsolike is a component of the eggshell.** *Curr Biol* 2003, **13:**1058-1063.
153. Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B: **RFECS: a random-forest based algorithm for enhancer identification from chromatin state.** *PLoS Comput Biol* 2013, **9:**e1002968.
154. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.** *Nat Genet* 2007, **39:**311-318.
155. Kang K, Kim J, Chung JH, Lee D: **Decoding the genome with an integrative analysis tool: combinatorial CRM Decoder.** *Nucleic Acids Res* 2011, **39:**e116.
156. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.** *Genome Res* 2005, **15:**1034-1050.
157. Goldman M, Craft B, Swatloski T, Cline M, Morozova O, Diekhans M, Haussler D, Zhu J: **The UCSC Cancer Genomics Browser: update 2015.** *Nucleic Acids Res* 2014.
158. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al: **The developmental transcriptome of *Drosophila* melanogaster.** *Nature* 2011, **471:**473-479.
159. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, Eswara P, O'Connor MJ, Schwartz S, Miller W, Chiaromonte F: **Distinguishing regulatory DNA from neutral sites.** *Genome Res* 2003, **13:**64-72.
160. Yakoby N, Bristow CA, Gong D, Schafer X, Lembong J, Zartman JJ, Halfon MS, Schupbach T, Shvartsman SY: **A combinatorial code for pattern formation in *Drosophila* oogenesis.** *Dev Cell* 2008, **15:**725-737.
161. Janssens H, Hou S, Jaeger J, Kim AR, Myasnikova E, Sharp D, Reinitz J: **Quantitative and predictive model of transcriptional control of the *Drosophila* melanogaster even skipped gene.** *Nat Genet* 2006, **38:**1159-1165.

162. Ashyraliyev M, Siggens K, Janssens H, Blom J, Akam M, Jaeger J: **Gene circuit analysis of the terminal gap gene huckebein.** *PLoS Comput Biol* 2009, **5:**e1000548.
163. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB: **Large-scale turnover of functional transcription factor binding sites in Drosophila.** *PLoS Comput Biol* 2006, **2:**e130.
164. Maston GA, Landt SG, Snyder M, Green MR: **Characterization of enhancer function from genome-wide analyses.** *Annu Rev Genomics Hum Genet* 2012, **13:**29-57.
165. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV: **OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs.** *Nucleic Acids Res* 2013, **41:**D358-365.
166. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al: **Pfam: the protein families database.** *Nucleic Acids Res* 2014, **42:**D222-230.
167. Barrett T, Edgar R: **Gene expression omnibus: microarray data storage, submission, retrieval, and analysis.** *Methods Enzymol* 2006, **411:**352-369.
168. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31:**248-250.
169. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33:**D514-517.
170. Wang T, Gu J, Li Y: **Inferring the perturbed microRNA regulatory networks from gene expression data using a network propagation based method.** *BMC Bioinformatics* 2014, **15:**255.
171. Jacquemin T, Jiang R: **Walking on a tissue-specific disease-protein-complex heterogeneous network for the discovery of disease-related protein complexes.** *Biomed Res Int* 2013, **2013:**732650.
172. Reimand J, Tooming L, Peterson H, Adler P, Vilo J: **GraphWeb: mining heterogeneous biological networks for gene modules with functional significance.** *Nucleic Acids Res* 2008, **36:**W452-459.
173. Shen R, Goonesekere NC, Guda C: **Mining functional subgraphs from cancer protein-protein interaction networks.** *BMC Syst Biol* 2012, **6 Suppl 3:**S2.
174. Liu Y, Gu Q, Hou JP, Han J, Ma J: **A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression.** *BMC Bioinformatics* 2014, **15:**37.
175. Mostafavi S, Morris Q: **Combining many interaction networks to predict gene function and analyze gene lists.** *Proteomics* 2012, **12:**1687-1696.
176. Ivan G, Grolmusz V: **When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks.** *Bioinformatics* 2011, **27:**405-407.
177. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
178. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res* 2011, **39:**W29-37.
179. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al: **Ensembl 2015.** *Nucleic Acids Res* 2014.
180. Tong H, Faloutsos C, Pan J-Y: **Fast Random Walk with Restart and Its Applications.** In *Proceedings of the Sixth International Conference on Data Mining*. pp. 613-622: IEEE Computer Society; 2006:613-622.

181. Glenn AL, Yang Y: **The potential role of the striatum in antisocial behavior and psychopathy.** *Biol Psychiatry* 2012, **72:**817-822.
182. Kucharski R, Maleszka R, Hayward DC, Ball EE: **A royal jelly protein is expressed in a subset of Kenyon cells in the mushroom bodies of the honey bee brain.** *Naturwissenschaften* 1998, **85:**343-346.
183. Drapeau MD, Albert S, Kucharski R, Prusko C, Maleszka R: **Evolution of the Yellow/Major Royal Jelly Protein family and the emergence of social behavior in honey bees.** *Genome Res* 2006, **16:**1385-1394.
184. Parker R, Guarna MM, Melathopoulos AP, Moon KM, White R, Huxter E, Pernal SF, Foster LJ: **Correlation of proteome-wide changes with social immunity behaviors provides insight into resistance to the parasitic mite, Varroa destructor, in the honey bee (Apis mellifera).** *Genome Biol* 2012, **13:**R81.
185. Dean J, Ghemawat S: **MapReduce: simplified data processing on large clusters.** *Commun ACM* 2008, **51:**107-113.
186. Kang U, Tsourakakis CE, Faloutsos C: **PEGASUS: A Peta-Scale Graph Mining System Implementation and Observations.** In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*. pp. 229-238: IEEE Computer Society; 2009:229-238.
187. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB: **PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.** *Nat Biotechnol* 2009, **27:**66-75.
188. Low Y, Bickson D, Gonzalez J, Guestrin C, Kyrola A, Hellerstein JM: **Distributed GraphLab: a framework for machine learning and data mining in the cloud.** *Proc VLDB Endow* 2012, **5:**716-727.
189. Xin RS, Gonzalez JE, Franklin MJ, Stoica I: **GraphX: a resilient distributed graph system on Spark.** In *First International Workshop on Graph Data Management Experiences and Systems*. pp. 1-6. New York, New York: ACM; 2013:1-6.
190. Koller D, Friedman N: *Probabilistic graphical models : principles and techniques.* Cambridge, MA: MIT Press; 2009.
191. Kang U, Chau DH, Faloutsos C: **Mining large graphs: Algorithms, inference, and discoveries.** In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*. pp. 243-254: IEEE Computer Society; 2011:243-254.
192. Yu X, Ren X, Sun Y, Gu Q, Sturt B, Khandelwal U, Norick B, Han J: **Personalized entity recommendation: a heterogeneous information network approach.** In *Proceedings of the 7th ACM international conference on Web search and data mining*. pp. 283-292. New York, New York, USA: ACM; 2014:283-292.