The Spatial and Temporal Regulatory Code of Transcription Initiation in *Drosophila melanogaster*

by

Elizabeth Ann Rach

Computational Biology and Bioinformatics
Duke University

Date:_____

Approved:

_____
Uwe Ohler, PhD, Supervisor

_____
Fred S. Dietrich, PhD

_____
Jack Keene, PhD

_____
Sayan Mukherjee, PhD

_____
Gregory Crawford, PhD

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in Computational Biology and Bioinformatics
in the Graduate School of Duke University

2010

ABSTRACT

The Spatial and Temporal Regulatory Code of Transcription Initiation in *Drosophila melanogaster*

by

Elizabeth Ann Rach

Computational Biology and Bioinformatics
Duke University

Date:_____

Approved:

_____
Uwe Ohler, PhD, Supervisor

_____
Fred S. Dietrich, PhD

_____
Jack Keene, PhD

_____
Sayan Mukherjee, PhD

_____
Gregory Crawford, PhD

An abstract of a dissertation submitted in partial fulfillment
of the requirements for the degree of Doctor of Philosophy
in Computational Biology and Bioinformatics in the
Graduate School of Duke University

2010

Abstract

Transcription initiation is a key component in the regulation of gene expression. Recent high-throughput sequencing techniques have enhanced our understanding of mammalian transcription by revealing narrow and broad patterns of transcription start sites (TSSs). Transcription initiation is central to the determination of condition specificity, as distinct repertoires of transcription factors (TFs) that assist in the recruitment of the RNA polymerase II to the DNA are present under different conditions. However, our understanding of the presence and spatiotemporal architecture of the promoter patterns in the fruit fly remains in its infancy. Nucleosome organization and transcription initiation have been considered hallmarks of gene expression, but their cooperative regulation is also not yet understood.

In this work, we applied a hierarchical clustering strategy on available 5′ expressed sequence tags (ESTs), and developed an improved paired-end sequencing strategy to explore the transcription initiation landscape of the *D.melanogaster* genome. We distinguished three initiation patterns: “peaked or Narrow Peak TSSs”, “Broad Peak TSSs”, and “broad TSS cluster groups or Weak Peak TSSs”. The promoters of peaked TSSs contained the location specific sequence elements, and were bound by TATA Binding Protein (TBP), while the promoters of broad TSS cluster groups were associated

with non-location-specific elements, and were bound by the TATA-box related Factor 2 (TRF2).

Available ESTs and a tiling array time series enabled us to show that TSSs had distinct associations to conditions, and temporal patterns of embryonic activity differed across the majority of alternative promoters. Peaked promoters had an association to maternally inherited transcripts, and broad TSS cluster group promoters were more highly associated to zygotic utilization. The paired-end sequencing strategy identified a large number of 5′ capped transcripts originating from coding exons that were unlikely the result of alternative TSSs, but rather the product of post-transcriptional modifications.

We applied an innovative search program called FREE to embryo, head, and testes specific core promoter sequences and identified 123 motifs: 16 novel and 107 supported by other motif sources. Motifs in the embryo specific core promoters were found at location hotspots from the TSS. A family of oligos was discovered that matched the Pause Button motif that is associated with RNA pol II stalling.

Lastly, we analyzed nucleosome organization, chromatin structure, and insulators across the three promoter patterns in the fruit fly and human genomes. The WP promoters showed higher associations with H2A.Z, DNase Hypersensitivity Sites (DHS), H3K4 methylations, and Class I insulators CTCF/BEAF32/CP190. Conversely, NP

promoters had higher associations with polII and GAF binding. BP promoters exhibited a combination of features from both promoter patterns. Our study provides a comprehensive map of initiation sites and the conditions under which they are utilized in *D. melanogaster*. The presence of promoter specific histone replacements, chromatin modifications, and insulator elements support the existence of two divergent strategies of transcriptional regulation in higher eukaryotes. Together, these data illustrate the complex regulatory code of transcription initiation.

Dedication

To my wonderfully strong and loving parents, Kathy and Herb.

To the world you might be one person,

but to one person, you just might be the world. *– unknown author*

Contents

List of Tables

List of Figures

List of Abbreviations

A - adenine

ARTS – Accurate Recognition of Transcription Starts

BAP/TAP – bacterial alkaline phosphatase/tobacco acid pyrophosphatase

BDGC/BDGP – Berkeley *Drosophila* Genome Collection (Project)

BEST – Binding site Estimation Suite of Tools

bgm - bubblegum

bp – base pairs

BP – Broad Peak

C - cytosine

CAGE – Capped Analysis of Gene Expression

cDNA – complementary DNA

CDS – coding sequence

ChIP – chromatin immunoprecipitation

CoA – coenzyme A

CRF – conditional random field

CTCF – CCCTC binding factor

DHS – DNAse Hypersensitivity Site

DNA – Deoxyribose Nucleic Acid

DPE – downstream core promoter element

DPN – Depleted Proximal Nucleosome

DRE - DNA replication element

EM – expectation maximization

EPD – Eukaryotic Promoter Database

ER – endoplasmic reticulum

EST – Expressed Sequence Tag

eve – *even-skipped*

G – guanine

GAF – GAGA Associated Factor

GHMM – generalized HMM

GO – Gene Ontology

H(D)AT – Histone (De)Acetyl Transferase

HMM – hidden Markov model

iid – identically independently distributed

IMM – interpolated Markov Model

INR – initiator

kb – kilobase

L-BFGS – Limited memory Broyden Fletcher Goldfarb Shanno

MAP – maximum a posteriori

MCMC – Markov Chain Monte Carlo

me - methylation

MM – Markov model

mRNA – mature RNA

MLF – motif location function

MTE – motif ten element

NCBI – National Center for Biotechnology Information

NFR – nucleosome free region

NIEHS – National Institute of Environmental Health Sciences

NP – Narrow Peak

NURF – NUcleosome Remodeling Factor

OPN – Occupied Proximal Nucleosome

PC – Pearson correlation

PCR – polymerase chain reaction

PEAT – Paired End Analysis of TSSs

per - period

PIC – pre-initiation complex

PSSM – position-specific scoring matrix

PWM – position weight matrix

RACE – Rapid Amplification of cDNA Ends

RCA – rolling circle amplification

RNA – Ribo Nucleic Acid

RNA pol II – RNA polymerase II

RT-PCR – reverse transcription PCR

SAGE – Serial Analysis of Gene Expression

ssRNA – single stranded RNA

SVM – support vector machine

T - thymine

TAF – TBP-associated factor

TBP – TATA-box binding protein

TF – transcription factor

TFBS – TF binding site

tim - timeless

TRF2 – TBP-related factor 2

TSS – Transcription Start Site

ttk – tramtrack

U – uracil

UTR – UnTranslated Region

VLCFA – very-long-chain-fatty-acids

WAM – weight array matrix

WP – Weak Peak

YMF – Yeast Motif Finder

Acknowledgements

I would like to acknowledge the academic support of my advisor Uwe Ohler, and committee members Fred Dietrich, Jack Keene, Sayan Mukherjee, and Greg Crawford. While it was never official, Fred Dietrich acted as a co-advisor throughout my PhD, and I am very grateful to the discussions we've had, and the guidance he's provided. I am most especially thankful for Fred letting me use the office in his lab during the last 2 years of my PhD, and for the many volleyball games we've played during the summer. I would also like to thank my high school math teacher Mr. George Chilmonik for his inspiration and motivation. He showed me that life has no boundaries, and if you work hard enough, you can achieve your dreams. Without him, I never thought I'd be where I am today.

I am most grateful to my family for being a strong emotional support network. My mom and Dad have always been there for me, in thick and thin. Without them, none of this would have been possible. My Aunts Marion and Patricia have opened their homes and hearts, my Uncle Michael and Aunt Heidi have shared many hours on the phone, and my Godmother Laurie has never forgotten to send care boxes on a birthday or holiday. To all of the Rachs and Hanes who have provided words of encouragement, and to Keri, my best friend of 14 years.

Most of all, I'd like to thank God for patiently guiding and helping me over the years. He has been with me in the hospital when I have suffered from psoriatic arthritis, and He has been with me at the height of my academic pursuits. I have always felt His intimate presence, and for that I have become very fond of this poem:

Footprints in the Sand

One night I dreamed I was walking along the beach with the Lord.

Many scenes from my life flashed across the sky.

In each scene I notice footprints in the Sand.

Sometimes there were two sets of footprints,

Other times there was only one.

This bothered me because I noticed that during the low periods of my life,

when I was suffering from anguish, sorrow or defeat,

I could see only one set of footprints.

So I said to the Lord

"You promised me Lord, that if I followed you,

You would walk with me always.

But I have noticed that during the most trying periods of my life,

There has been only one set of footprints in the sand.

Why, when I needed you most, have you not been there for me?"

The Lord replied, "The times when you have seen only one set of footprints,

My child, is when I carried you."

~Mary Stevenson (http://www.footprints-inthe-sand.com)

1. The DNA Code For Gene Regulation

1.1 From Sequencing to Regulation

Since its initial discovery by Watson and Crick in 1953, Deoxyribose Nucleic Acid (DNA) has transformed science and spawned the genome era (Watson and Crick 1953). DNA is composed of four bases: adenine (A), thymine (T), guanine (G), and cytosine (C) that are joined together millions of times in a double helix. While the occurrence of these four bases can be random, in many regions, the sequence is organized in a precise way to give the cell directions on how to function and interact with its environment. The challenge of the genome era has been to distinguish meaningful sections of the genome from those that provide less information to the cell, and to decipher the relationship between the cryptic DNA code and the biochemical functions that it regulates.

In the 1860s, Gregor Mendel studied the phenotypic traits of pea plants and conceived of the notion of a basic unit of inherited information, which was later termed a 'gene' by Wilhelm Johannsen (Churchill 1974). In the 20th century, the term 'gene' came to be used to characterize stretches of DNA holding the cells information, and a 'genome' was used to note all of the genes within one organism. Through the study of genes, or genomics, it was found that DNA is fundamental to all six kingdoms of life (eubacteria, archaebacteria, protista, fungi, plantae, and animalia). Due to the complexity

of the DNA, and the lack of methods to analyze the sequence of a genome in a high throughput manner, genes were initially studied individually. With the advancement of a range of technologies in the late 20th century, entire genomes of small eukaryotes were sequenced for the first time (The yeast genome directory 1997). Sequencing efforts escalated with public and private efforts in competition to finish the human genome, which was annotated in 2001 (Lander et al. 2001).

With access to the DNA code of an entire organism, researchers believed they would be able to understand the relationship between genes, their expression, and their functional consequences within the cell. However, this proved to be much more challenging than initially thought because DNA is not processed into an expressed protein directly. Instead, DNA is processed into an intermediate form known as RiboNucleic Acid (RNA), which is then translated into a protein that is expressed in the cell. RNA is single stranded and is composed of adenine (A), cytosine (C), guanine (G), and Uracil (U). One region of DNA can produce multiple isoforms of a mature RNA (mRNA) can be generated from the same gene, each having a slightly different sequence. This indirect coupling between gene and protein was further complicated by the discovery of additional regulatory elements, including microRNAs, that alter the levels of mRNA abundance, and repertoires of mRNA composition over time and in different tissues (Bartel 2004). Thus, the sequencing of DNA did not solve the gene-expression-

function problem, but rather, it introduced a plethora of additional challenges to be addressed.

1.2 Drosophila melanogaster

1.2.1 An Ideal Model Organism

Advancements in genetics have been made primarily possible through the use of model organisms, or living models used to understand genomic phenomenon. While all living things contain DNA, and therefore can be used as model organisms, certain species have properties that make them more opportune than others. For instance, organisms with a short life span enable scientists to study the passing of genes from one generation to the next in a timely fashion. A smaller genome enables scientists to more easily isolate features with experimental techniques, such as mutagenesis, or changing the DNA. With a 10-day life span and a compact genome, the common fruit fly *Drosophila melanogaster* has both of these attributes. In addition, it has a high fecundity, laying up to 100 eggs per day, and males and females can be easily distinguished by their morphological segmentation (Sang 2001). *D.melanogaster* is small, and easy to grow in the lab, making it cost efficient and an ideal organism for scientists.

D.melanogaster has been successfully used to advance genetics since the late 1800s when Charles Woodworth initially suggested the fruit fly as a model organism (Nobel

lectures physiology or medicine 1922-1941 1965). The fruit fly is most widely recognized for its extensive use in the early 1900s in demonstrating how genes are organized into chromosomes (Nobel lectures physiology or medicine 1922-1941 1965). More recently, the fruit fly's remarkably similar physiology has been used to further genetic advancements in higher eukaryotes. Fruit flies possess body parts, such as eyes, and legs, and also an immune and nervous system. Studies on behavior showed that *Drosophila* responded to external stimuli in a corresponding fashion to humans. When *Drosophila* were exposed to high levels of alcohol, the ability of their motor functions decreased until they passed out, resembling signs of human intoxication (Morozova, Anholt, and Mackay 2006). During their intoxication, expression changes were measured for 582 genes, giving scientists clues into the genetic basis of alcohol impairment. Biochemical responses observed in *Drosophila* have also been used to gain insight into genetic diseases. For instance, mutations, or changes, to the *Drosophila troponin I* gene, and transgenic fruit fly strains containing a mutant form of human delta-sarcoglycan *deltasg(S151A)* were shown to have modified cardiac chambers with impaired heart functioning. This has enabled the use of *Drosophila* to study human cardiomyopathy (Wolf et al. 2006).

1.2.2 Embryonic Development

Perhaps, the most beneficial use of *Drosophila* has been to gain insight into the development of an organism from an egg to an adult. Upon fertilization of a *Drosophila* egg by a sperm, an embryo is created that undergoes morphological changes marked by developmental stages during the first 24 hours. The embryo hatches into a series of three larval stages that are separated by molts during which morphological changes continue. After five days, the third instar larva undergoes pupation and remains a pupa until metamorphosis four days later. After nine days, the fly emerges as an adult, equipped with a head, mouth, eyes, antennae, three thoracic segments, and eight or nine abdominal segments. The segments are uniquely identifiable, but share a similar plan for pairs of legs, wings, and small knob-shaped structures that provide balance during flight, called halteres (The development of drosophila melanogaster 1993).

The organization of body segments is established early during embryogenesis. Like other insects, but unlike mammals, immediately after fertilization, a series of nuclear divisions occur without cellular divisions. The nuclei migrate to the outside of the embryo where the cell boundaries begin to form, while the precursors to eggs or sperm called pole cells, migrate to the posterior of the egg. At this stage, the embryo is called a cellular blastoderm. Before it reaches this phase, the embryo relies on maternal effects: mRNA and proteins that were in the cell before fertilization or deposited from

germline nurse cells. Once it has reached the cellular blastoderm stage, the embryo initiates its own machinery to provide the nutrients needed during development in the maternal to zygotic transition (Benoit et al. 2009). Scientists have been able to capture the existence of cascades of expression profiles for transcription factors and developmentally regulated genes before and after the cellular blastoderm phase. Using *in situ* hybridization, the Berkeley *Drosophila* Genome Project (BDGP) has created a large database of over 75,000 images of dynamic gene expression profiles at different stages during *Drosophila* embryogenesis (see Figure 1) (Tomancak et al. 2002). These images have given great insight into the spatial patterning of gene expression and their dynamic transformations over time. While the genetics of human development differs from *Drosophila melanogaster*, the fruit fly has provided a wealth of information about common genomic principles that are shared across both organisms.

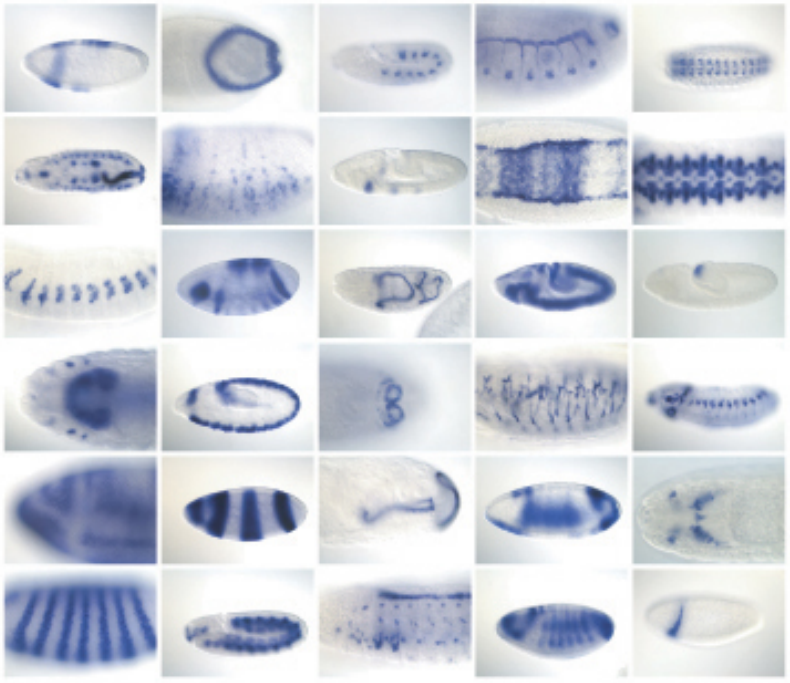


Figure 1: Collection of *In situ* Images

Each image depicts the spatial expression of a gene during a stage of embryogenesis (Tomancak et al. 2002).

1.2.3 Availability of 12 Genomes

Drosophila continues to remain an important model organism for the future of genetics. In 2007, a consortium completed the sequencing of 12 *Drosophila* genomes (see Figure 2) (Clark et al. 2007). This has enabled researchers to analyze the evolution of genes across species over time. By identifying gene duplications, losses, and gains, scientists are able to derive information about the functions of genes and their importance for the survival of the species. Concurrent efforts are underway to capture

the repertoires of RNA, and other regulatory factors in the 12 *Drosophila* species (modENCODE 2010). With this information, even more precise findings can be made into the complex evolutionary processes that produce RNA and regulate the expression of genes. The 12 *Drosophila* genomes distinguish the fly from other model organisms in being the largest phylogeny of available genomic data for any one genus to date. Much like the transition from sequencing to regulation, the work on *Drosophila melanogaster* in this thesis provides a foundation for future explorations in the 11 sister *Drosophila* species.

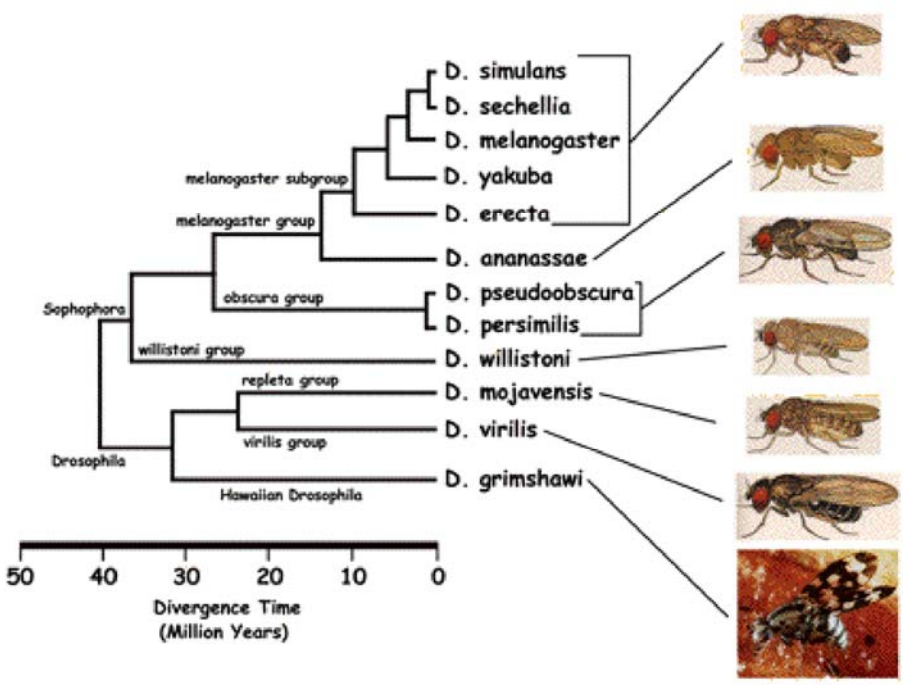


Figure 2: Phylogeny of 12 Sequenced *Drosophila* Genomes

The 12 fly genomes are organized into a phylogeny based on the similarity of their DNA sequence. The divergence time denotes the millions of years ago from the present time during which unique phenotypic features and genotypic sequence evolved for a species. Species with less DNA shared across genomes are more distantly related to each other than those that diverged recently and have a high similarity of sequence (Camos and Badia 2006). Subgroups and groups denote the common ancestors of *Drosophila* species.

1.3 Genetics of Transcription

1.3.1 DNA is Tightly Compacted Within a Cell

Transcription is an essential regulatory process in the expression of genes and the proper development and fundamental maintenance of an organism. Before we can understand the genetic components of gene expression, we must first address the structure of the DNA, and the alterations needed to make the DNA ready for transcription, replication, and other processes. The *Drosophila melanogaster* genome consists of ~137 million base pairs (bp), and the human genome contains ~3.2 billion bp. As both the fruit fly and human are diploid organisms, there are ~274 million bp of DNA in each fruit fly cell and ~6.4 billion bp in each human cell. Each bp of DNA is .34 nanometer (1 billionth of a meter), resulting in $(.34 \times 10^{-9}) \times (.274 \times 10^{9}) = .09$ meters of DNA in each fruit fly cell, and $(.34 \times 10^{-9}) \times (6.4 \times 10^{9}) = 2.2$ meters of DNA in each human cell. As there are ~50 trillion cells in the human body, this total ~100 trillion meters of DNA per human. The distance from the Earth to the sun is 150 billion meters, which means each of us has enough DNA to go from here to the sun more than 300

times (Annunziato 2008)! With dimensions of this size, the DNA must undergo various levels of condensation to fit within the small space of each cell's nucleus (see Figure 3).

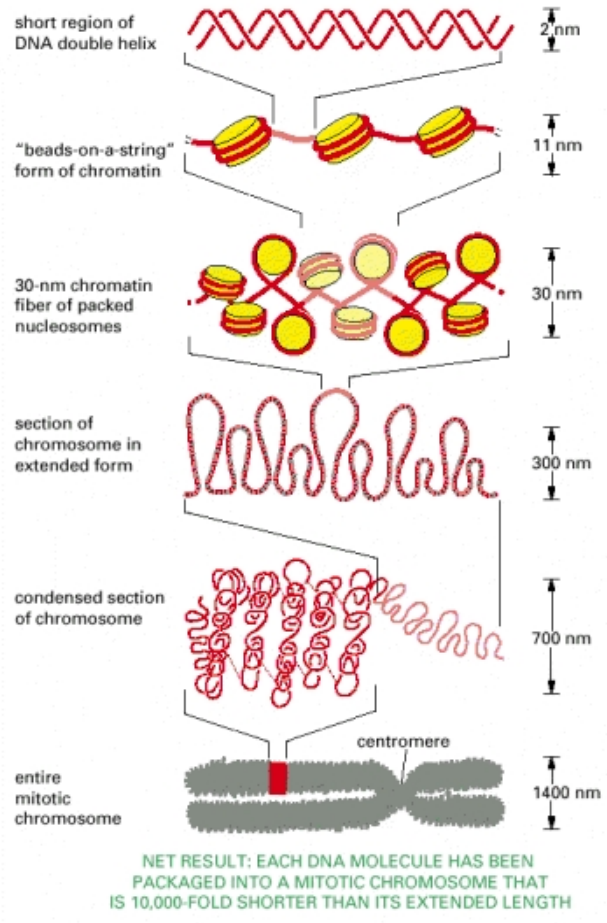


Figure 3: Compact Chromatin Structure

DNA (red) is wrapped around histones (yellow) and must undergo various levels of condensation before it can be properly packed into the nucleus of a cell. The same packing strategies are conserved across species, in spite of differences in the DNA sequence (Alberts et al. 2002).

The DNA begins as a double stranded helix that is wound around an octomer of four core histones (H2A, H2B, H3, H4) to form a nucleosome. The 146bp of DNA in one nucleosome is locked in place by a linker histone (H1), and the assembly of multiple nucleosomes across the DNA resembles beads on a string. The nucleosomes are packaged into 30nm chromatin fibers that are condensed into loop domains. A tertiary structure of loop domains provide further condensation for the DNA, as protein scaffolds establish the final shape of a chromosome. Because humans and *Drosophila* are both diploid organisms, one set of chromosomes is inherited from the mother, while another is inherited from the father. In humans, there are 22 pairs of autosomes and one pair of sex chromosomes, while *Drosophila* consists of three pairs of autosomes (2,3,4), and one pair of sex chromosomes (X/Y). Chromosomes 4 and Y are small in size and contain very few genes (Celniker et al. 2002).

1.3.2 Epigenetic Modifications Increase DNA Accessibility

The compact organization of the DNA limits its accessibility to the binding of the transcriptional machinery. As a result, epigenetic modifications to the histones, chromatin, and nucleosomes are required by activator proteins and other regulatory complexes to open up the DNA. N-terminal tails extend from each of the four core histones (H2A, H2B, H3, H4) within a nucleosome and can be modified. Changes in the

histones alter the stability of the secondary structure of the chromatin and the 30nm fiber. In general, the methylation of lysines, acetylation of lysines, and phosphorlyation of serines on the histone tails signify the destabilization of the chromatin, while the ubiquitination of lysines restablizes the chromatin. Enzymes, such as the Histone Acetyl Transferases (HAT), and the Histone DeAcetyl Transferases (HDAT), perform these modifications. As more than one modification may occur on each N-terminal tail, the assortment of possible combinations of methylations, acetylations, phosphorlyations, and ubiquitinations create the histone code (Jenuwein and Allis 2001).

Modifications can also be made to the loops of DNA surrounding the histones to produce the effect of chromatin destabilization. The ISW2 and SWI/SNF modifying complexes in yeast destabilize the interactions between the histones and DNA through an ATP dependent manner (Zofall et al. 2006). ISW2 and SWI/SNF cause the DNA to disassociate from the histone's edge, and form a displaced bulging loop 9-11bp and 50bp increments, respectively. The loops are moved along the surface of the histone in a wave-like manner, and DNA rebinds to the histone on the opposing side of the wave, resulting in no alterations to the total number of histone-DNA contacts (Zofall et al. 2006).

Nucleosome remodeling is the last type of modification that can destabilize the chromatin and open it up for transcription. The Nucleosome Remodeling Factor

(NURF) functions by sliding the entire spherical nucleosome package along the DNA in 10bp increments (Schwanbeck, Xiao, and Wu 2004). NURF uses an ISWI ATPase pump to catalyze this reaction, and has been shown to regulate genes involved in *Drosophila* innate immunity (Kwon et al. 2008). NURF synergistically facilitates transcription with the modifications of histones (Mizuguchi et al. 2001), and variants of NURF components have been shown to contain histone binding specificities (Kwon et al. 2009).

Once the chromatin is destabilized and open stretches of DNA are made available, regulatory complexes and activating proteins bind to the DNA and interact with each other to further modify the chromatin structure (see Figure 4). Regulatory complexes can bind directly upstream of a gene, or be located several kilobases (kb) away. They can have either a high specificity and bind to a region of DNA upstream of a single gene, or a low specificity and bind at various locations throughout the genome. Once the regulatory complexes are in place, the activating proteins provide the final epigenetic modifications necessary for transcription.

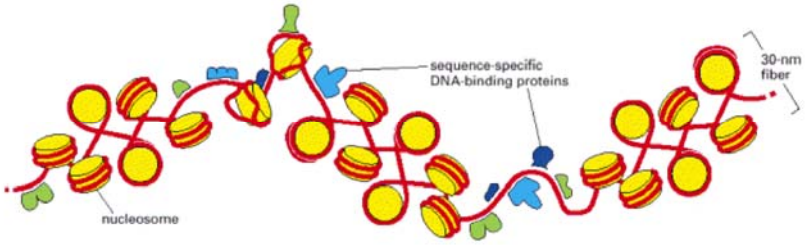


Figure 4: Chromatin Primed for Transcription

A 30nm chromatin fiber is ready for transcription after regions of the chromatin (red) have been destabilized around the octomer of histones (yellow). Activator and regulatory proteins (green, blue, navy) bind to the open stretches of DNA, making the final epigenetic modifications and priming the DNA for transcription (Alberts et al. 2002).

1.3.3 Initiation

After epigenetic modifications have made the DNA accessible, transcription factors (TFs) are recruited to bind to short motifs in the DNA (see Figure 5). A gene can contain multiple binding sites upstream in the promoter, or downstream of the most 3' end, and even within it. Factors that affect the transcription of the gene and are encoded at it's locus are called cis-regulatory elements, while those encoded at a different locus, such as transcription factors, are called trans-factors (Latchman 2005).

One model of transcription initiation involves the well-studied general transcription factor TFIID, which consists of the TATA-box Binding Protein (TBP) and 10-14 TBP-Associated Factors (TAFs) that bind approximately 30bp upstream of the 5' end of the gene to the TATA box, and to other sequence motifs in the core promoter (Latchman 2005). This enables the binding of the general transcription factors TFIIB, TFIIA, TFIIF, TFIIE, and TFIIH that help recruit a complex consisting of the RNA polymerase II, the mediator, and over 100 proteins, called the RNA polymerase holoenzyme, to the site of transcription initiation in the DNA (see Figure 5) (Latchman 2005). In *Drosophila melanogaster*, the exact site of transcription initiation is marked by the

Initiator (INR) sequence motif (Lo and Smale 1996). Although the RNA polymerase II enzyme does not have a direct affinity for the DNA, with the guidance of the general TFs, it is responsible for transcription of the majority of eukaryotic genes (protein-coding genes and many regulatory RNAs). Together, these proteins assemble into the pre-initiation complex (PIC) (Smale and Kadonaga 2003).

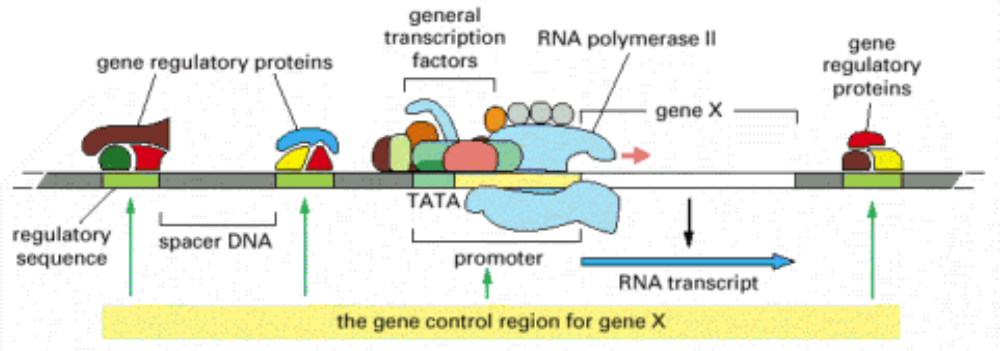


Figure 5: Transcription Initiation

Regulatory protein complexes and transcription factors bind to DNA regulatory sequence elements and recruit the RNA polymerase II to the DNA (Alberts et al. 2002).

1.3.4 Elongation and Termination

Upon the phosphorlyation of its C-terminal domain tail, the RNA pol II holoenzyme complex is released from the grip of the cis-regulatory factors, and transcriptional elongation begins. Transcription proceeds for approximately 20-30 bases and pauses until the C-terminal domain is phosphorlyated for a second time (Latchman

2005). As transcription continues, the RNA pol II complex reads each base of DNA. A temporary DNA-RNA hybrid is formed during complementary base pairing that is quickly separated, to restore the double DNA helix, and the single strand of the newly synthesized RNA molecule (Latchman 2005). During pairing, the DNA bases A,T,G,C align with the RNA bases U,A,C,G, respectively. When the entire length of the gene has been read, the RNA pol II complex terminates transcription by disassociating from the DNA. Studies have shown that for some genes, after termination, TFIIF remains associated with TFIIA and TFIID to allow for repeated cycles of transcription (Latchman 2005).

1.3.5 Post Transcriptional Processing

The resulting RNA transcript goes through a series of processing steps before it is expressed as a protein. Regions of the DNA are selectively included (exons) and excluded (introns) from the transcript in a process called splicing. Splicing can initiate simultaneously with elongation and continue throughout transcriptional termination (Gunderson and Johnson 2009). A 5′ cap and 3′ poly(A) tail are added to the transcript to prevent it from being degraded resulting in a mature messenger RNA (mRNA). The mRNA binds to distinct export proteins that transport it from the nucleus into the cytoplasm (Latchman 2005). If it is not destroyed or stored in p-bodies (Aizer and Shav-

Tal 2008), the mRNA binds to ribosomes, the site of protein synthesis. Ribosomes can occur freely in the cytoplasm, but a large proportion of them are located on the membrane of the rough endoplasmic reticulum (ER). In the ER, the mRNA is translated into a protein, properly folded, and released into the cytoplasm. Improperly translated, or folded proteins may be recognized and destroyed. Failure to do so can be detrimental to the organism and ultimately decrease functioning and lead to diseases. Proteins with similar functions localize together in distinct areas of the cell (Pyhtila et al. 2008).

1.4 Transcriptional Regulators

1.4.1 Enhancers and Repressors

The rate of transcription is precisely regulated by sections of DNA that regulatory complexes, activator proteins, and transcription factors can bind. This binding can increase or decrease the rate of transcription. DNA sections that increase the rate of transcription when bound are called enhancers, while those that decrease the rate of transcription are repressors. Multiple enhancers can synergistically increase the rate of transcription higher than would result from the sum of each enhancer individually. Enhancers and repressors can function independently or work in conjunction with each other. An activator protein may bind to an enhancer upstream of the promoter region to assist in the recruitment of individual transcription factors to the DNA. A repressor

protein may inhibit this transcriptional activation by competitively binding to an adjacent DNA repressor sequence (see Figure 6A), or by obstructing the structural domain intended for the transcription factor (see Figure 6B). The DNA of an enhancer motif may fold back or undergo structural changes that allow its activator protein to interact with the RNA polymerase II. A repressor sequence can inhibit this by the direct binding of its repressor protein to the RNA pol II. This may readjust the alignment of the enhancer sequence and move the bound activator protein out of reach of the RNA pol II (see Figure 6C).

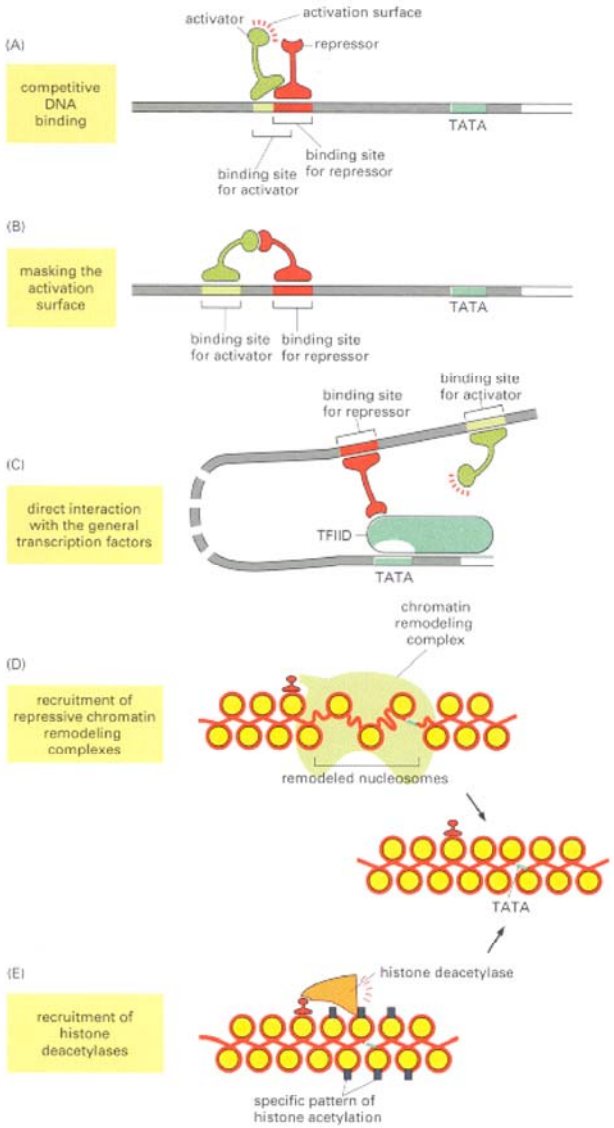


Figure 6: Models for Enhancer and Repressor Activity During Transcriptional Activation

Enhancers and repressors can affect the binding of individual TFs to promoter regions, TF interactions facilitated by DNA structure, and chromatin modifications (Alberts et al. 2002).

Enhancers and repressors can also affect the rate of transcription through epigenetic regulation. An activator protein bound to an enhancer may recruit a chromatin or nucleosome remodeling complex that opens up the DNA and makes it accessible to the transcriptional machinery. This can be blocked by the binding of a repressor at the exact location of remodeling, or by a repressor bound protein actively recruiting factors that inhibit the activity of the chromatin remodeling complex (see Figure 6D). In addition, activator proteins bound to enhancers can recruit enzymes, such as acetylases, that modify histone tails and open up the chromatin for transcription initiation. Similarly, repressor bound proteins can recruit enzymes, such as deacetylases, that remove the histone tail modifications and limit the accessibility of the DNA (see Figure 6E).

A single gene can respond to multiple cues from a combination of enhancers and repressors in the DNA. One of the best examples of this is the regulation of the gene *even-skipped (eve)* in *D.melanogaster*. Clusters of enhancer and repressor sequence elements are found upstream of the gene. During embryogenesis, activating and repressing proteins bind to combinations of these regulatory elements to generate seven stripes of *eve* expression. Studies have shown that stripe two occurs only where the activator proteins *bicoid* and *hunchback* are present, and the repressor proteins *giant* and *kruppel* are absent (Small et al. 1991). This demonstrates the sheer selectivity and

complexity of enhancer and repressor regulation that is possible in the *Drosophila* genome.

1.4.2 Operons and Insulators

An operon is a group of genes that is controlled by one promoter. An operator (either an enhancer or a repressor) interacts with the promoter to create a polycistronic transcript. Genes in an operon typically have the same transcriptional orientation and function, and the intergenic distance between them is small. There are three main types of operons (Blumenthal 2004). In Type I, one polycistronic mature RNA with one poly(A) tail is produced from which multiple proteins are translated (Figure 7A). Type I operons are most commonly found in bacteria and archae. For Type II operons, one polycistronic pre-mRNA is generated from which separate monocistronic mRNAs are processed, each with their own poly(A) tail (Figure 7B). Nematodes contain high occurrences of Type II operons. In Type III operons, one polycistronic pre-mRNA is produced from one promoter, and only one type of mRNA is generated (Figure 7C). Type II and Type III operons are not commonly thought of as conventional operons because the mRNAs are not polycistronic (Blumenthal 2004).

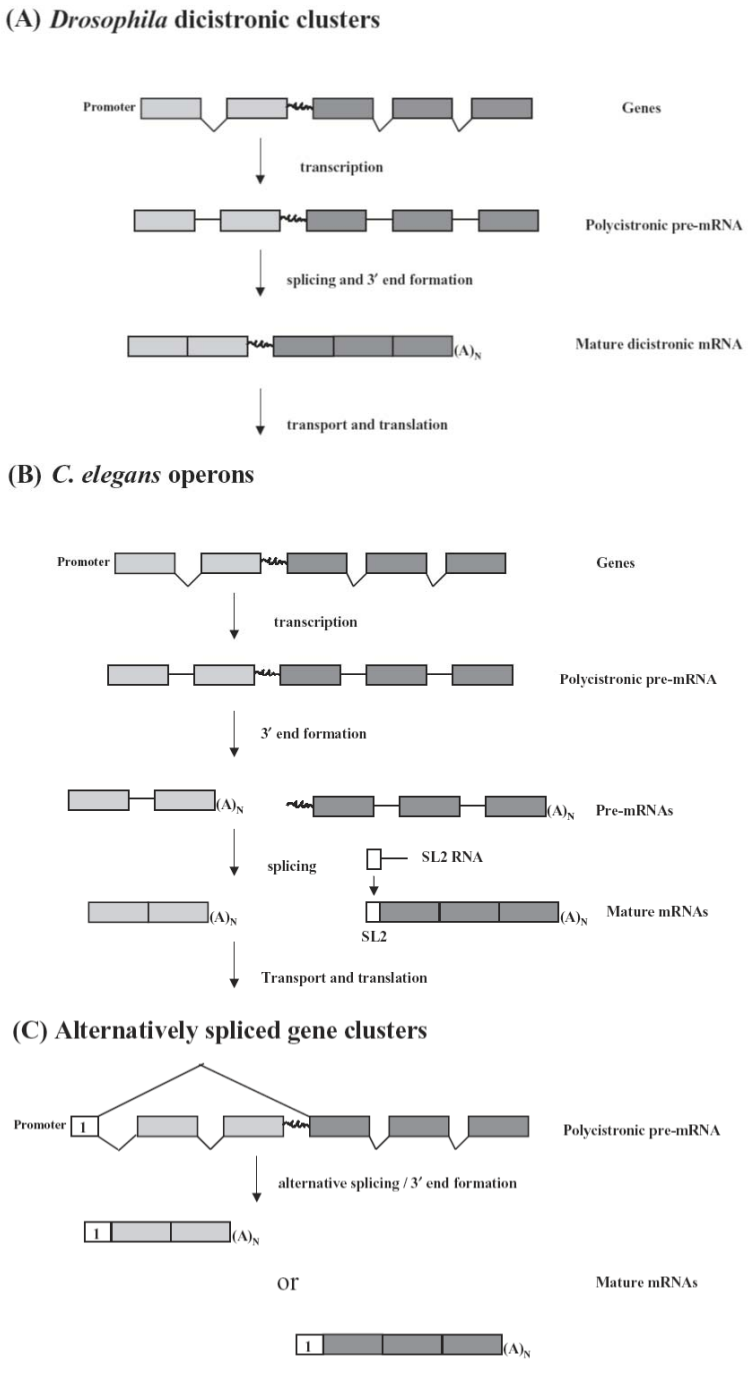


Figure 7: Three Types of Operons

(A) Type I operons generate mature dicistronic transcripts from a polycistronic pre-mRNA. **(B)** Type II operons produce mature monocistronic transcripts from one polycistronic pre-mRNA, through the trans-splicing SL2 RNA. **(C)** Type III operons in which one type of monocistronic mRNA is processed from a polycistronic pre-mRNA (Blumenthal 2004).

While operons are not prevalent in higher eukaryotes, dicistronic clusters (Type I) are present throughout the *D.melanogaster* genome (Misra et al. 2002). Specific examples include the *Adh* related genes that are involved in alcohol processing (Brogna and Ashburner 1997), *stoned A* and *stoned B* that localize to nerve terminals (Andrews et al. 1996), and *mei-217* and *mei-218* that govern meiotic recombination (Liu et al. 2000). Type II operons have not been found in the fruit fly, and only one instance of a Type III operon has been identified with the ubiquitin-ribosomal protein fusion (Mottus et al. 1997).

Insulator motifs function differently from operons, as they isolate the effects of transcription to individual genes or domains. The presence of an insulator sequence between two adjacent genes can promote vastly different levels of expression and specificities of cell types in which they are expressed. In *D.melanogaster*, dCTCF is a well-studied insulator element that often occurs between closely positioned promoters (Smith et al. 2009). Insulators can also serve as boundary elements to prevent the spread of heterochromatin (Schedl and Broach 2003).

1.5 Condition Specific Transcriptional Programs

The complexity of transcription increases dramatically when one considers that the concentration and repertoire of regulatory complexes, activator proteins, and transcription factors varies across cell types and at different time periods. This is most apparent in the asymmetrical localization of protein combinations during cell divisions, as illustrated in figure 8. Initially, the parent cell begins with a set of regulatory proteins, designated by the brown irregular shaped oval. A new regulatory protein is generated and the cell divides, resulting in one daughter cell with the regulatory protein, and the second daughter cell without it. Two more regulatory proteins are generated and each daughter cell divides, resulting in four cells with either one or two regulatory proteins. At this stage, cells may also exist with none, or all three regulatory proteins, but are not depicted here. The process is repeated with the generation of two additional proteins, and the division of the granddaughter cells. This produces eight embryonic cells with different combinations of regulatory proteins.

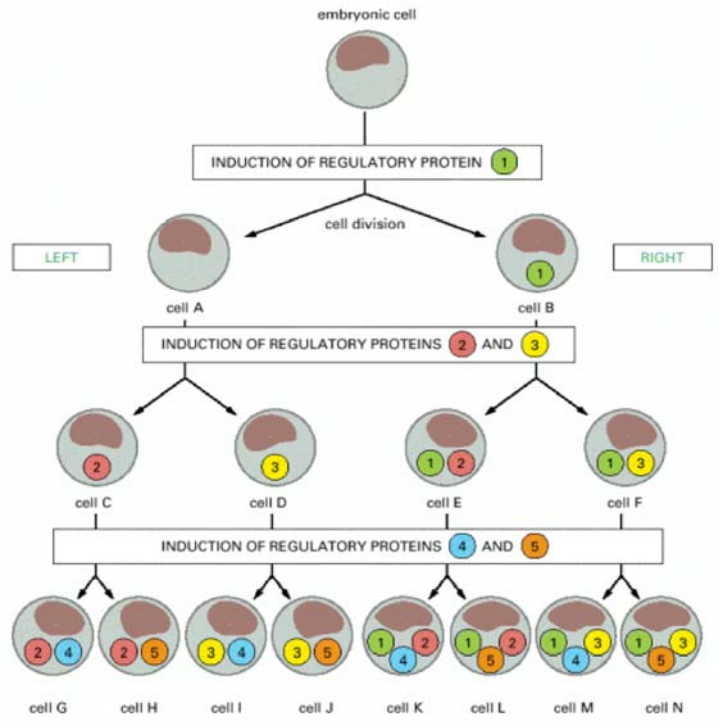


Figure 8: Diverse Combinations of Regulatory Proteins Exist Across Cells

For illustrative purposes, all of the cells are depicted with the same orientation (see 'Left', 'Right') (Alberts et al. 2002).

If we consider each generation of cells a time period, we can see how the diversity of regulatory proteins increases throughout temporal development. Similarly, if we characterize the function of each cell by the combination of proteins that regulate it, we can see how the repertoires of regulatory factors change across cells with divergent functions. The determination of function is a difficult challenge facing scientists today. To assess the problem, collections of cells with similar functions are often grouped together as 'cell types'. In *D.melanogaster*, tissues and organs are

commonly used for spatial comparisons, and the morphological stages of embryogenesis are typically used in temporal analyses. The type of spatial or temporal association is referred to as the 'condition.'

The simple model of combinatorial dynamics in Figure 8 can be directly applied to transcription. The regulatory proteins may depict transcription factors whose specific concentrations are spatiotemporal signatures of expression profiles. One instance of this in *D.melanogaster* is the existence of testes specific TAFs in the basal transcriptional machinery (Metcalf and Wassarman 2007). The combinations of regulatory proteins may also reflect binding sites that are enhancers, repressors, and insulators. The gene *yellow* is regulated across *Drosophila* species by different enchancers that produce pattern specific expression profiles. Some species have enhancers that generate black spots on their wings, while others have lost enhancers that control the pigmentation of their abdomen (see Figure 9) (Unicellular organisms: Genomes 2010). The combinatorial nature of the regulatory proteins may further reflect the assortment of epigenetic modifications around promoter regions that provide genomic landmarks for the transcriptional machinery. However, this is only beginning to be explored in *D.melanogaster*.

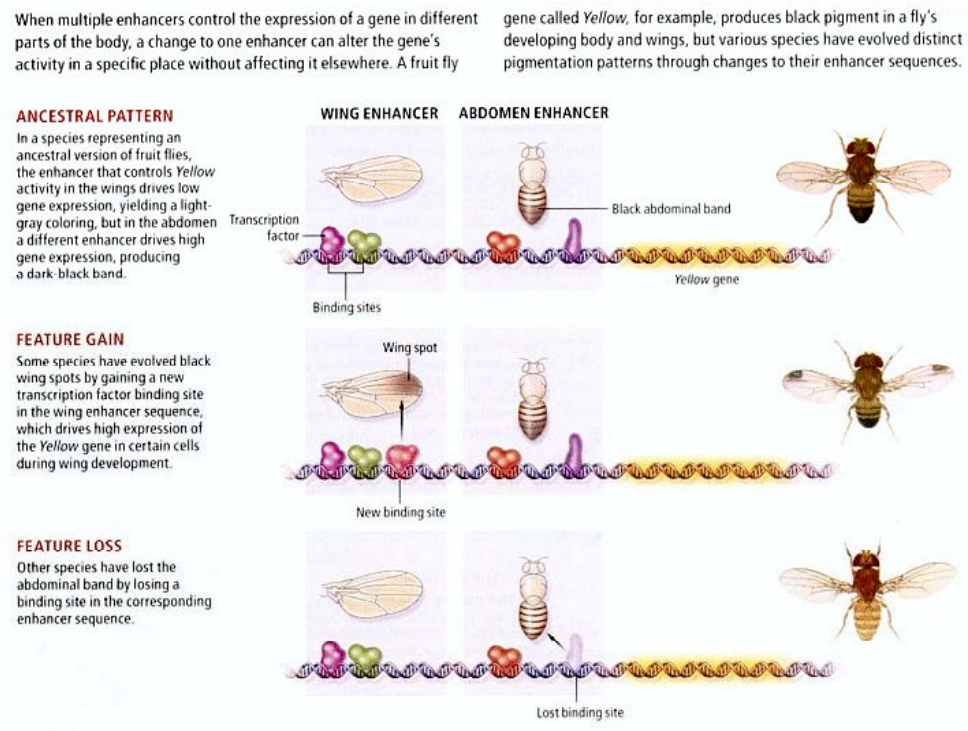


Figure 9: Combinations of Enhancers Produce Condition Specific Expression of the Gene *Yellow* (Unicellular organisms: Genomes 2010)

1.6 Associations of Transcription to Disease

Disruptions to condition specific transcriptional programs can have serious effects on proper functioning, and may lead to diseases. For instance, the gene *bubblegum (bgm),* located on chromosome 2L, is actively transcribed during *Drosophila* embryogenesis and is responsible for ligating very-long-chain-fatty-acids (VLCFA) with coenzyme A (CoA) into fatty aceyl-CoA (Min and Benzer 1999). This is the first reaction in the complete breakdown of fatty acids for energy. VLCFAs are highly concentrated in

the brain, and changes to the transcription of this gene result in a build up of VLCFAs in the optic lobes, characterized by bubbly neurodegeneration (see Figure 10). These effects can be treated tby administering the dietary supplement of GTO-glycerol trioleate oil at specific stages of development (see Figure 10). A corresponding disease is found in humans called Adrenoleukodystrophy (ALD). ALD is x-linked, and can also be treated by supplementing Lorenzo's oil into the diet (Min and Benzer 1999). This shows that spatiotemporal specific transcriptional programs are not only important for proper functioning, but they can also impact the delivery of treatments.

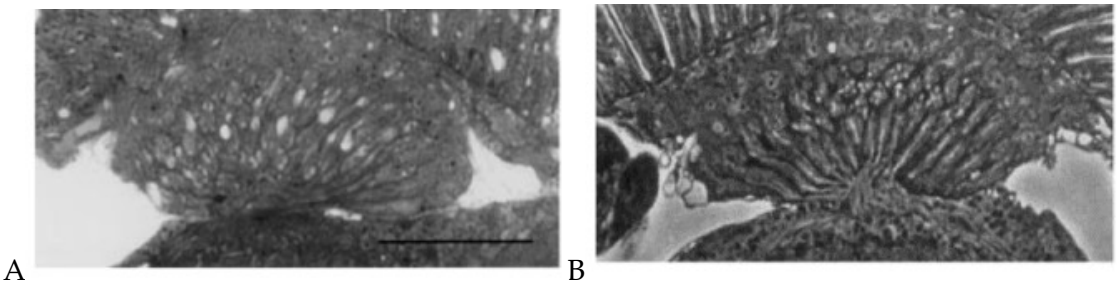


Figure 10: Treatment of *bubblegum* Neurodegeneration is Time Sensitive

(A) The optic lobes of an adult mutant bgm male that had GTO in its diets for 15 days still had bubbly degeneration on the lamina. Larva that was raised on GTO medium showed little damage **(B)** (Min and Benzer 1999).

2. Experimental and Computational Strategies for Modeling 5' Ends

The starting locations of translation can be identified by the presence of the AUG start codon throughout organisms. Mapping starting locations of transcription is a much more challenging task because no such universal codon exists. Historically, TSSs were mapped using various low throughput technologies for individual genes. More recently, the advancement of technology has promoted high throughput methods that provide TSS data for many genes. Progress has also been made in the computational arena, as efforts have moved from using DNA sequence to identify TSS locations and/or promoter motifs, to using high throughput data to characterize promoter properties genome wide. Each experimental and computational method has its advantages and disadvantages. Until all of these issues are addressed, better strategies are needed to assess TSS locations and promoter architecture.

2.1 Low Throughput Experimental Methods

2.1.1 S1 Mapping

Initial attempts to experimentally capture 5' ends of transcripts used the S1 mapping technology. This was performed by designing a probe that mapped to a genomic region of interest. The probe was hybridized to RNA or single stranded DNA, and S1 nuclease was used to degrade single stranded regions not bound by the probe. The product was run on a gel, and the resulting size of the band and sequence of the

probe was used to determine the TSS (see Figure 11) (National-Diagnostics S1 mapping 2010). The TSSs for the genes encoding Ca(2+)-ATPase and ADP-ribosyl cyclase in *Drosophila*, whose function is important in the ATP pathway for energy production, were mapped using this methodology in conjunction with primer extension (see section 2.1.3) (Magyar, Bakos, and Varadi 1995; Nata et al. 1995).

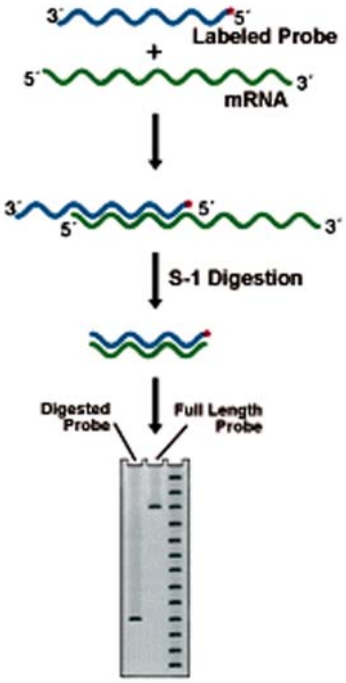


Figure 11: Protocol for S1 Mapping (National-Diagnostics S1 mapping 2010)

An advantage of S1 mapping was that more than one probe could be evaluated at the same time, provided that the bands produced on the gel were of different sizes. The biggest disadvantage of S1 mapping was that the general location of the TSS had to

be known a priori. If the probe could mapped to downstream sequence, the experiment had to be tediously reproduced using probes further upstream. In addition, mismatches in DNA:RNA binding could be resistant to nuclease degradation, and if the probe mapped to an intron-exon boundary, or completely inside an intron, a band would not be produced, resulting in an uninformative experiment.

2.1.2 RNase Protection

The S1 mapping technology was slightly improved upon with the development of RNase protection. In this technique, labeled RNA probes were designed to bind entirely within the coding region. Then, an excessive quantity of the probe was mixed with the RNA sample, and RNase, instead of a general nuclease, was used to digest single stranded RNA. The product was run on a gel, and the size of the band and sequence of the probe were used to determine the TSS. The amount of probe protected from digestion was measured after it was run on a gel by autoradiography. This reflected the amount of transcribed RNA (see Figure 12) (National-Diagnostics Ribonuclease protection 2010). The TSSs of the *period* (*per*) and *timeless* (*tim*) genes that encode components of the circadian rhythm in *Drosophila*, were determined using RNase protection (Cheng, Gvakharia, and Hardin 1998).

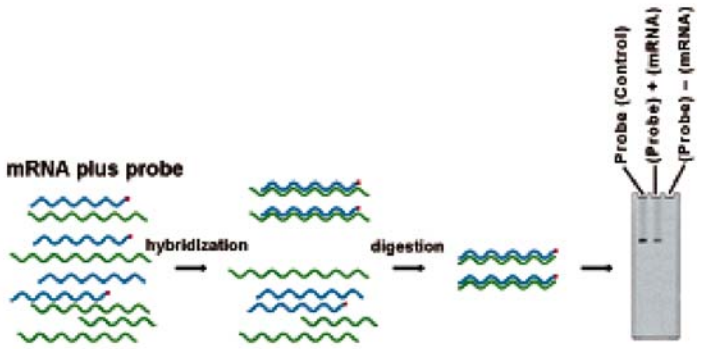


Figure 12: Protocol for RNase Protection (National-Diagnostics Ribonuclease protection 2010)

With this assay, the quantity of RNA could be measured, and a higher sensitivity of RNA capture was achieved. Similar to S1 mapping, multiple probes could be tested for different transcripts at the same time, as long as different sized bands were produced. A major disadvantage of RNase protection was that the general location of the TSS had to be known a priori, and this often required tedious repition of the experiment. While probes mapping to intron-exon boundaries were no longer a problem, if your probe hybridized completely inside of an intron, a band would still not be evident.

2.1.3 Primer Extension

The next strategy that was developed to capture the 5' ends of transcripts was primer extension. In this method, a probe was designed to bind to the internal region of a specific gene. Then, reverse transcriptase was used to extend the sequence upstream

of the probe, creating a strand of complementary DNA (cDNA). The product was run on a gel, and the size of the band was used to determine the TSS. The cDNA could be sequenced to more precisely define the intron/exon sequence of the transcript. If an excess of probe was used, the quantity of RNA transcribed could be measured (see Figure 13) (National-Diagnostics Primer extension 2010). In *D.melanogaster*, the TSS for the *E2F* gene was mapped with this strategy and its promoter was shown to contain the transcriptional regulatory motif called the DNA replication-related element (DRE) (Sawado et al. 1998).

The advantage of using primer extension was that introns, or intron-exon boundaries were not a problem. Also, Polymerase Chain Reaction (PCR) was not used to more accurately assess the quantity of RNA transcripts. As a result, it could only be performed on highly expressed genes. Like previous methods, the general area of the TSS had to be known, so a probe was designed relatively close to the 5′ end. In addition, the probe had to be specifically designed to avoid mapping to the 5′ end of another transcript generated from the same gene. A disadvantage of primer extension was that it had to be repeated, as the reverse transcriptase frequently fell prematurely still remained a problem.

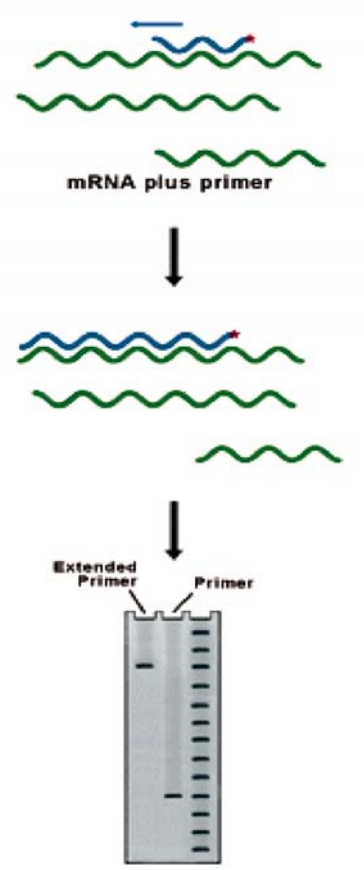


Figure 13: Primer Extension (National-Diagnostics Primer extension 2010)

2.1.4 5' RACE

5′ Rapid Amplification of cDNA Ends (5′ RACE) laid the foundation for the development of future high-throughput techniques. There were two protocols for performing 5′ RACE. The first relied on the fact that mRNAs had poly(A) tails. A poly dT primer was bound to the poly(A) tails, and a reverse transcriptase was used to extend the cDNA sequences to the 5′ end of the transcripts. Upon reaching the 5′ end, the reverse transcriptase left a CCC overhang to which a second universal primer

containing a GGG sequence would bind. The reverse transcriptase switched templates and extended transcription through the universal GGG oligo. In theory, this produced complete cDNAs for all of the transcrips that extended from the poly(A) tails to the oligo capped 5′ ends. Then, for the analysis of individual genes, specific primers were used in PCR to amplify the 5′ region of interest. The product was run on a gel, and sequenced to identify the TSS (see Figure 14) (Biosciences 2003). 5′ RACE was successfully used in combination with primer extension (see section 2.1.3) to map the TSS of the *parkin* gene in *D.melanogaster* (Bae, Park, and Kang 2003). *Parkin* is a component of the *Drosophila* model of Parkinson's disease (Venderova et al. 2009).

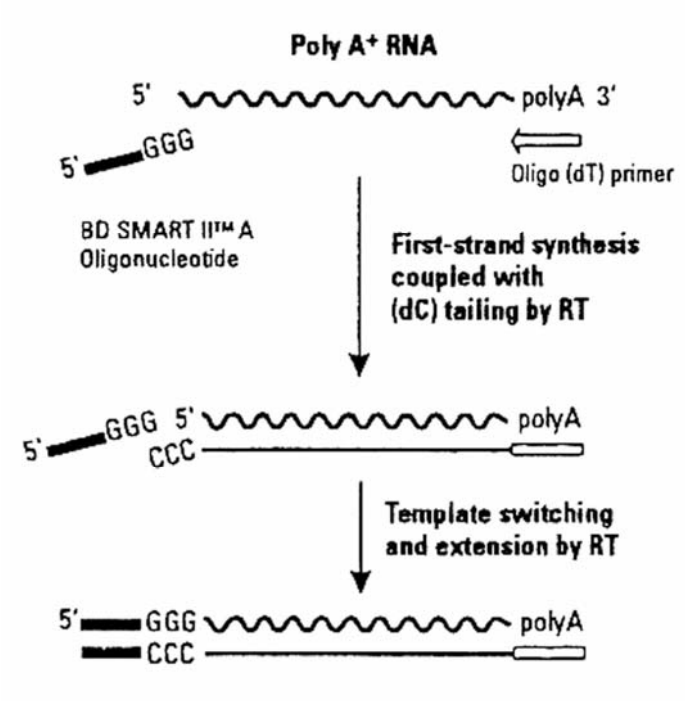


Figure 14: 5' RACE (Biosciences 2003)

The advantage of the first 5′ RACE protocol was that once the pool of complete transcripts was generated for all mRNAs, it could be allocated to various 5′ RACE experiments targeting different genes. In addition, 5′ ends could be mapped for genes that were not highly expressed, and the general location of the TSS did not have to be known *a priori*. Like primer extension, introns and intron-exon boundaries did not pose problems.

There were two main issues when using the first protocol of 5′ RACE. The first was that when using a poly-T primer, the reverse transcriptase would often fall off the mRNA template before reaching the most 5′ end, which resulted in incomplete transcripts. This was exacerbated by the second problem of PCR bias, in which shorter sequence fragments are amplified faster than longer sequence fragments. As a result, more copies of the shorter fragment were produced, lending support to the false notion that the shorter fragment was the true transcript, while the longer fragment resulted from experimental error.

To address these issues, a modified second 5′ RACE protocol could be implemented in which a gene-specific primer close to the TSS was used instead of a poly-T primer. Then, the sequence was reverse transcribed akin to that of the first protocol, using a universal template switching primer (Strachan and Read 1999). This modified protocol achieved a higher rate of capturing the 5′ ends of complete transcripts, however, PCR bias still resulted from fragments of different size. Because

only one transcript could be evaluated at a time, the pool of complete mRNAs was not available for each experiment. Furthermore, annealing temperatures had to be adjusted precisely.

2.2 High Throughput Experimental Methods

2.2.1 Expressed Sequence Tags (ESTs)

The development of Expressed Sequence Tags (ESTs) was the first implementation of high throughput technology to map parts of transcripts that could be used to annotate TSSs. To generate ESTs, cells from a condition or tissue, such as the brain, were isolated so all of their mRNA could be extracted. A poly-T primer was bound at the 3′ end of each transcript, and reverse transcriptase was used to extend the sequence to the 5′ end of the transcript, making a cDNA:RNA hybrid. The RNA strand of the hybrid was degraded using RNase, and replaced with DNA nucleotides by DNA polymerase. The double-stranded cDNA was sequenced and mapped back to the genome (see Figure 15).

The gene annotations in Flybase (Wilson, Goodman, and Strelets 2008) were established from the Berkeley *Drosophila* Genome Collection (BDGC) (Celniker et al. 2002). This collection contains thousands of ESTs mapped in various conditions. While TSS mapping has improved over the years, the BDGC remains the largest collection of ESTs in fruit fly to date, and a valuable resource for understanding downstream mRNA

architecture. Collections of ESTs in *Drosophila* and higher eukaryotes have been deposited in dbEST as part of the National Center for Biotechnology Information (NCBI) database (Boguski, Lowe, and Tolstoshev 1993).

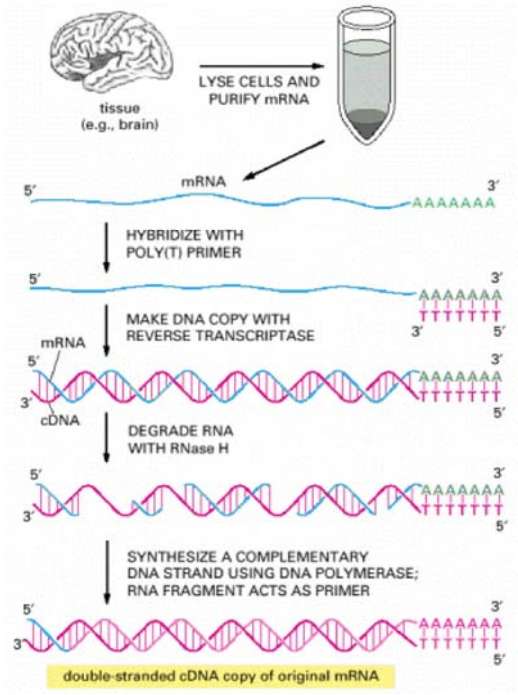


Figure 15: cDNAs

After cDNA libraries were generated, they were sequenced and aligned to the genome as ESTs (Alberts et al. 2002).

The main breakthough with ESTs was that data could be generated for multiple genes in the same experiment. In addition, EST fragments were 500-800bp in length, which enabled them to be mapped to the genome with high accuracy. They covered

multiple intron-exon boundaries, and gave insight into unique combinations of exons that were spliced out of transcripts.

Like earlier methods, the main disadvantage of ESTs was that the reverse transcriptase frequently fell off the template before completely transcribing to the most 5′ end of the transcript. Because PCR was not used, genes with low expression often had very low EST coverage. Also, it was time consuming and costly to make the many cDNA libraries that were needed to cover the whole transcriptome. As a result, RNA levels could not be accurately measured from ESTs; even genes with high expression sometimes had low coverage due to limited experimental resources.

2.2.2 Serial Analysis of Gene Expression (SAGE)

Serial Analysis of Gene Expression (SAGE) was initially developed as a contender for microarrays to quantify sequence expression. SAGE's popularity dwindled when the cost of microarrays was low, but increased greatly with the resurgence in sequencing. The SAGE protocol generated cDNAs for all of the transcripts akin to that for the ESTs. Poly(A) mRNA was extracted, and a poly-T probe was used as a primer. Reverse transcriptase was introduced to extend the sequence to the 5′ end of the transcript, and RNase was added to destroy the hybrid RNA. Lastly, DNA polymerase replaced the RNA nucleotides with DNA, creating double-stranded cDNAs (see Figures 15, 16) (Song and Wyse 2004).

Initial efforts implementing SAGE focused on capturing the 3′ ends of transcripts, and the 3′ ends of the cDNAs were bound to streptavidin-coated beads by their poly-T tails, leaving the 5′ ends available for modification. An anchoring enzyme that recognized specific 4bp sequences, such as GTAC, was used to cut the 5′ ends off of the cDNAs bound to beads. Since any 4bp sequence is common throughout the genome, nearly every cDNA was cut. Then, either linker A or linker B was ligated upstream of the GTAC sequences. Two linkers were used instead of one because their placement was used to determine the orientation of the transcripts after sequencing. Both linkers were designed to contain a restriction site and a tagging enzyme was introduced that recognized it and cut ~20bp downstream. This reduced the long cDNA products to short 20bp tags (see Figure 16) (Song and Wyse 2004).

Next, the 3′ ends of both products were ligated together into one ditag, and amplified using PCR. It was very unlikely that a pair of 5′ transcript ends would randomly segregate together more than once. Thus, the RNA level of each transcript could be reflected by counting the number of tag occurrences within unique ditag pairs. Then, the product was cut again using an anchoring enzyme to cleave off the linkers from either side of the ditags. This left sticky ends on the ditag fragments that bound to each other end-to-end, creating one long string of DNA. The concatenated DNA was sequenced and each fragment was aligned back to the genome (see Figure 16) (Song and Wyse 2004).

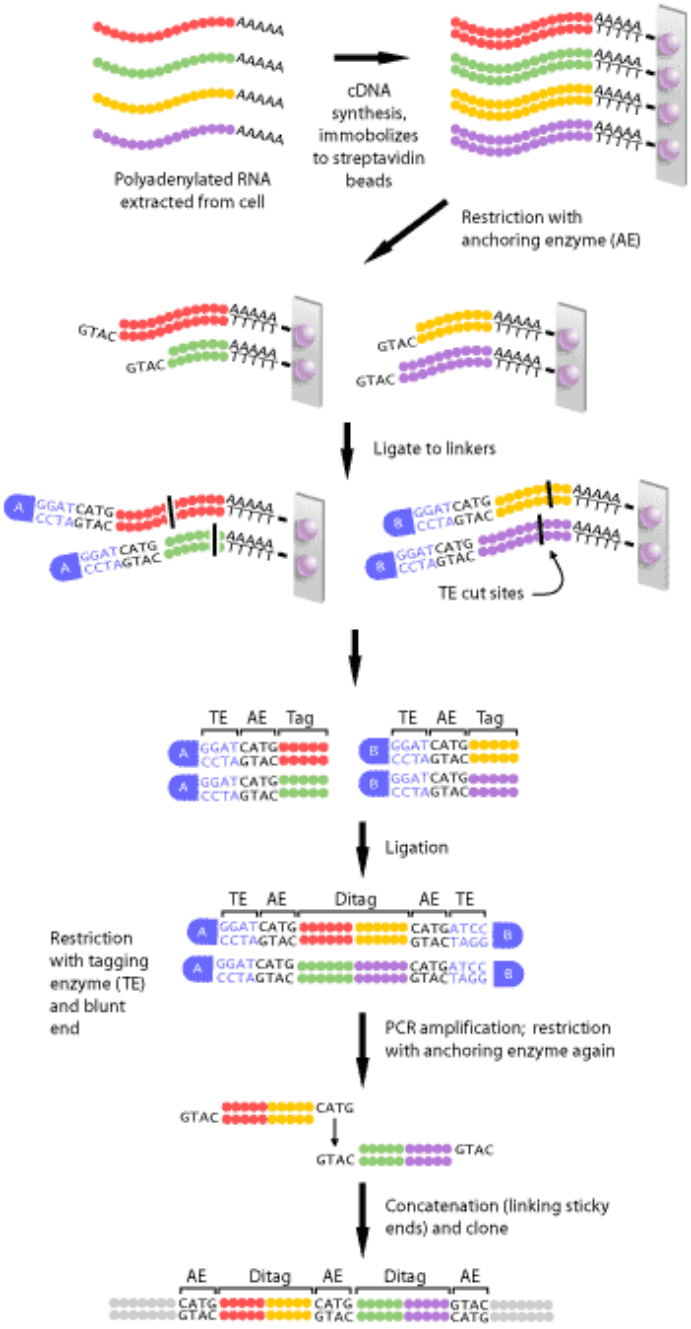


Figure 16: SAGE (Song and Wyse 2004)

One of the benefits of SAGE was that it could be used to rapidly identify the complete set of transcripts expressed in a certain condition or cellular process. While it did not map the complete architecture of every transcript, it provided a list of candidate genes that would have otherwise taken years to create. Genes with different transcript expression in male and female *Drosophila* heads and those involved in cell death were identified using SAGE (Fujii and Amrein 2002; Gorski et al. 2003).

By generating 20bp tags, the SAGE methodology was able to achieve high coverage of a multitude of transcripts in both a timely and cost-effective way. However, the 20bp tags also proved to be a downfall of the experiment, because their short sequences often aligned to multiple locations in the genome, especially for sequences with repetitive elements. Single or double sequencing or experimental errors also led to incorrect alignments to the genome. In cases of overlapping transcripts or genes, it was difficult to determine the exact origin of the tag. In addition, as the 4bp anchoring enzyme restriction site occurred regularly in sequence downstream of the TSS, the most 5′ end of transcripts was typically cut off. This lead to inaccurate calls of 5′ sites downstream of the true TSS locations. Even worse, if a certain mRNA did not have the enzyme recognition sequence, it could not be mapped at all. To alleviate some of these problems, a modified protocol was developed that used a template switching oligo to

more accurately map TSSs in *S.cerevisiae* (Zhang and Dietrich Mapping of transcription start sites in saccharomyces cerevisiae using 5' SAGE 2005).

2.2.3 Cap Analysis of Gene Expression (CAGE)

Cap Analysis of Gene Expression (CAGE) was invented by the biomedical research consortium RIKEN (Carninci et al. 2005). The protocol was similar to that of SAGE, with the exception of one fundamental difference. The 5′ ends of the transcripts were cap-trapped in order to ensure accurate mapping of the TSSs. Figure 17 shows how this was achieved by first selecting mRNA with poly(A) tails, and using random primers for reverse transcription. With random primers, reverse transcription could be initiated internally, rather than at the poly(A) 3′ tails. This decreased the distance the reverse transcriptase had to travel to the 5′ ends, and thus increased the occurrences of completely mapped transcripts. Although, the strategy was not perfect, since the reverse transcriptase was still able to prematurely fall off the RNA templates. To remove truncated cDNAs from the analysis, the cDNA:RNA hybrids were bound to streptavidin-coated beads by the properties of their 5′ caps. Then, RNase was introduced to destroy all single-stranded RNA (ssRNA). For full-length cDNAs, only ssRNA downstream of the random primer was destroyed. For truncated cDNAs, the ssRNA between the cap and the 5′ end of the cDNA was destroyed. This released the

cDNA:RNA hybrids from the streptavidin beads, and they were eliminated from the set of full-length cDNAs (Carninci et al. 1997).

The complete RNA:cDNA hybrids were recovered from the magnetic streptavidin beads, instead of cutting the 5′ ends as in SAGE, and a linker was added to the cDNA upstream of the 5′ ends, which preserved the exact locations of TSSs. Another RNase was used to destroy the RNA, and a double-stranded cDNA was created using DNA polymerase. The linker that was added was designed to contain a *Mme1* restriction site, and upon the addition of *Mme1*, 20bp tags were cut from the 5′ ends of each cDNA. To preserve the orientation of the transcripts, a second linker containing a different sequence was added to the 3′ of each tag, and the tags were amplified using PCR. Lastly, they were sequenced and aligned to the genome (see Figure 17) (Carninci et al. 2005).

RIKEN applied this technology in human and mouse to characterize promoters from tag frequencies and identify promoters selectively used in specific tissues. They also used CAGE to identify novel RNAs, and to model transcriptional dynamics using networks (see Figure 17) (Carninci et al. 2006). This technology has not yet been performed in *Drosophila*; however, various ESTs in the BDGC were generated from capped-trapped libraries (Stapleton et al. 2002).

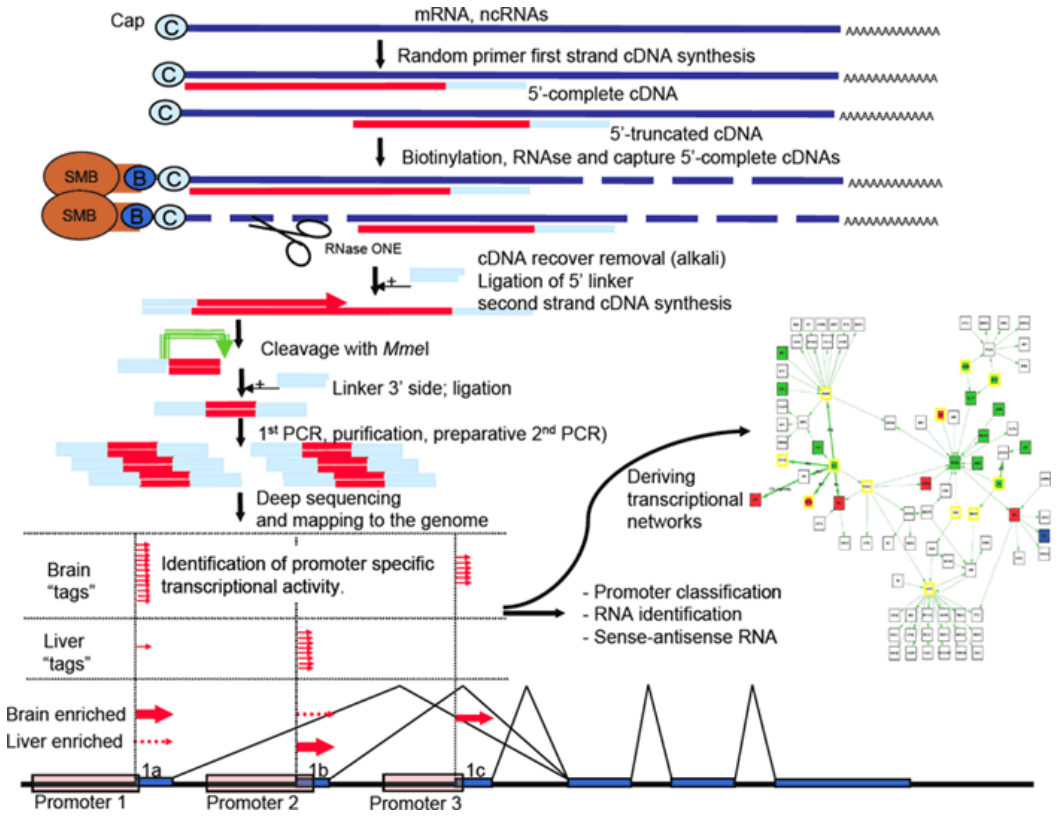


Figure 17: CAGE

This technology allowed the characterization of promoter architectures from tag frequencies, the evaluation of condition specific TSS utilization, the identification of novel RNAs, and the construction of transcriptional networks (RIKEN 2010).

CAGE was a major step towards the goal of accurately capturing the 5' ends of transcripts. This was a challenge that scientists had been facing for years. Because of its precision and the high magnitude of throughput, promoter locations could be more accurately identified, and RNA levels could be more precisely measured. Most importantly, CAGE advanced our knowledge of transcription initiation by showing that not only one TSS was utilized for each transcript, but rather broad and narrow patterns

of initiation spanned promoter regions. Because of their efforts, the study of promoters has made huge strides in a short period of time.

Like all 5' mapping technologies, CAGE too had its disadvantages. The issue of the 5' tags aligning to multiple or incorrect locations in the genome was not solved, nor had the problem of associating tags to overlapping genes or transcripts. In addition, because the mRNAs were not selected for size, tags could be generated from any RNA with a poly(A) tail that was transcribed by polII, including non-coding RNAs, such as microRNAs. As a confounding result of this problem, tags mapping to novel TSSs could not be directly linked to genes.

2.2.4 Tiling Arrays

Tiling arrays are a completely different strategy of measuring RNA expression than any of the high throughput methods previously discussed. Instead of a transcript-capture and sequencing-based approach, tiling arrays utilize relative differences in the hybridization of oligonucleotides across the genome. This method assumes the sequence of a genome is known, and uses it to detect new isoforms of transcripts.

To generate a tiling array, probes are designed that are ~25bp in length. Due to the vast size of a genome, typically millions of probes are designed that could be overlapping or non-overlapping. Overlapping probes provide higher levels of consistency in RNA expression over larger areas. Although, because they increase the

cost of the experiment, non-overlapping probes are often used (Manak et al. 2006). Next, a microarray experiment is performed at each of the million probes. In a microarray, DNA is sheared, and then, two cDNA probes are prepared, and fluorescently labeled with different colors. The first probe is the control probe for which the expression level is known. The second is the 25bp experimental probe with unknown expression. Aliquots of the DNA and both probes are mixed together, and a laser is used to excite the fluorescently labeled dyes upon the binding of their probe to the DNA (see Figure 18) (Gibson and Muse 2002).

The intensities of the green (control) dye and the red (experimental) dye are measured in the computer and compared to each other. If more of the experimental probe bound to the DNA, the well is recorded as red and that RNA is considered up-regulated. On the other hand, if more of the control probe bound to the DNA, the well is considered green and the RNA is considered down-regulated. Differences in magnitude of color are reflected by shades of red and green, and equal levels of experimental and control probe binding to the DNA is signified by a mixture of both colors that appears yellow. In Figure 18, black wells designate constitutive binding. Black is often used to signify a well in which the microarray experiment failed, and no binding was observed for either probe. The level of RNA expression of the experimental probe is digitally determined relative to the known level of expression of the control probe (see Figure 18) (Gibson and Muse 2002).

The tiling array experiment is repeated multiple times for every experimental probe, to more accurately assess RNA levels, and the normalized median values are often reported. This experiment was performed on the *Drosophila* genome every 2 hours during the first 24 hours of embryogenesis using the Affymetrix GeneChip2.0 tiling arrays (Manak et al. 2006), which has proven to be a valuable resource for studying development.

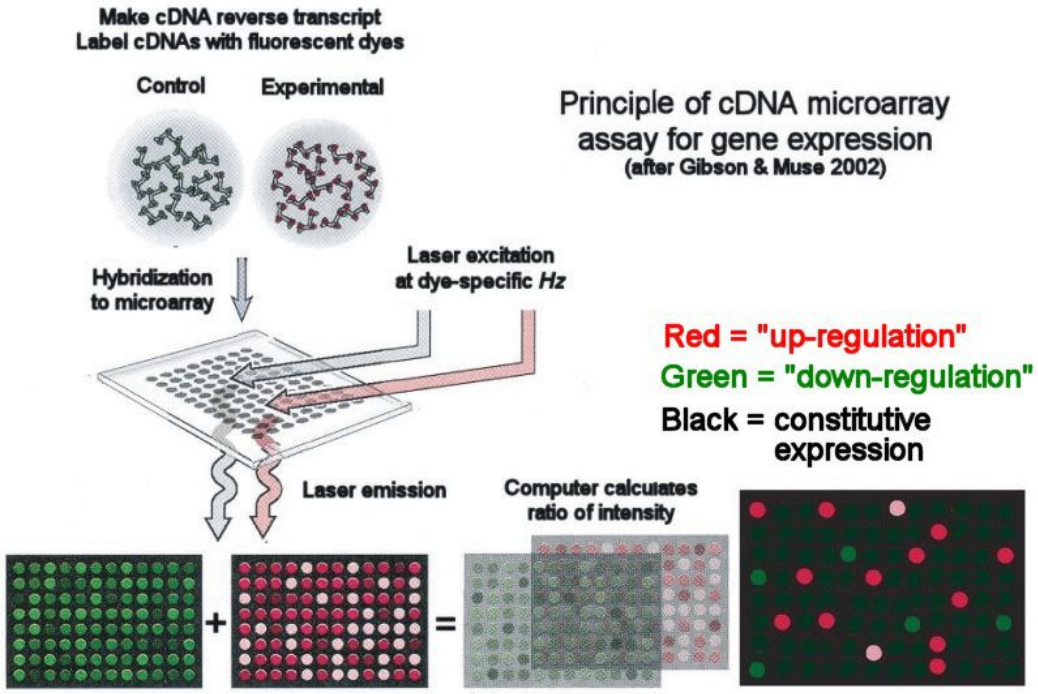


Figure 18: Microarray

A tiling array is created by performing individual microarray experiments across a genome (Gibson and Muse 2002).

Tiling arrays were developed using the well-established microarray technology, which gave scientists confidence in using them. The development of standard gene chips that could accommodate multiple microarray experiments made it easy for scientists to adopt as well. However, tiling arrays have a variety of pitfalls. The use of colors to determine RNA expression levels is not as precise as the counts of RNA transcripts found from the CAGE technology. The range of hybridization frequently gives a false color display for quantitative readout, with transcripts being assigned a fluorescence value that does not accurately reflect the true magnitude of their expression. In addition, some areas of the genome naturally hybridize to the oligos better than others, making the correlation of expression levels to colors imperfect. If the probes are not carefully designed, inaccurate measures of expression could also be produced from probes binding to multiple areas in the genome with 1 or 2 mismatches in the complementary sequence. These problems are confounded by probes mapping to intron-exon boundaries that produce inconsistent measures of fluorescence. This is especially detrimental when mapping TSSs.

2.3 Computational Identification of TSS Features

2.3.1 Classifying and Parsing

The annotation of a genome is the process of labeling the DNA based on its sequence features. The features provide meaningful biological insight into the utilization

of the genome. For example, two features may be introns and exons. Introns are removed from RNA during splicing, while exons are selectively maintained in mRNA and later translated into proteins. A genome is annotated after it is sequenced, as the function of the DNA is unknown. Annotating provides important information about the overall layout of the genome, and is essential for the identification of genes.

Annotation is performed through classifying and parsing. In the simple classification of input data, small windows of DNA are compared to sequence features using a probabilistic framework. The windows are labeled by their most probable feature. For instance, figure 19 shows that if the probability of a window is above a threshold, the label 1 is assigned; else, 0 is assigned. If a group of nucleotides in a stretch of DNA are divided according to their most probable feature, this constitutes parsing. Figure 19 shows that if the numbers 0,1,2,…,60 correspond to locations of nucleotides within a DNA sequence, one possible partition would be that nucleotides at locations 0-20 are intronic sequence (I), nucleotides at locations 21-55 are exonic sequence (E), and nucleotides at locations 56-60 are intergenic sequence (N).

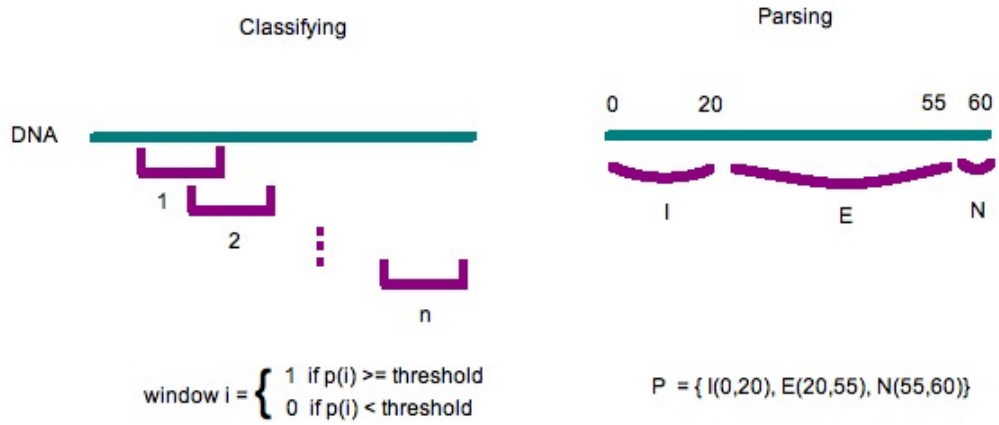


Figure 19: Genomic Annotation

For i = 1,…,n, window i can be classified as 1 or 0 based on the probability of the DNA belonging to a class of functional elements with respect to a threshold value. I (intron), E (exon), and N (intergenic sequence) denote one possible partition of DNA from locations 0 to 60.

In practice, classification and parsing have been used in more complex scenarios to computationally identify TSSs. For instance, in one study, the sequence upstream of TSSs was evaluated according to models of different combinations of frequently occurring core promoter motifs. Then, the parse of the sequence with the highest probability was determined from the classification of the models (Ohler 2006).

2.3.2 Sequence Classification

There are two types of classification techniques applied to input features $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$. In supervised training, the class labels $\mathbf{y} = (\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n)$ are known, and the

goal is to learn the mapping f: **x** -> **y**. In unsupervised learning, the class labels are not known, and **x** is partitioned into subsets $\mathbf{z} = \{(\mathbf{x}_2,\mathbf{x}_4),(\mathbf{x}_1,\mathbf{x}_3,\mathbf{x}_n),\ldots\}$ that are clustered to evaluate any underlying structure in the data.

In supervised training, an objective function is specified, and a model consisting of a classifier **f** is chosen. Then, the optimal parameters for the model are determined during training. There are two main types of classifiers: generative and discriminative. In a probabilistic framework, a generative classifier models the density of **x** in $p(\mathbf{x}|\mathbf{y})$, and new input variables **x** can be 'generated' by the class labels **y**. While, a discriminative classifier models the density of the labels **y** in $p(\mathbf{y}|\mathbf{x})$. No effort is put into modeling the input **x**, although, it can be obtained using Bayes Theorem:

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}$$

One of the caveats of applying a classifier is that the parameters may be learned so precisely from the training data that the model may not perform well on other data sets. This problem is called overfitting, and is avoided by applying a penalty to the weights of the feature functions, in a process called regularization.

The accuracy of the model can be evaluated using a procedure such as leave one out cross validation. In this method, the training data is divided into smaller sections, and the model is trained using all of the sections, except one. Then, the model is applied to the excluded section and because the class labels are known, the error between the model's predictions and the true class labels is found. This is repeated by excluding each

of the sections, and the error rates are averaged to give an overall error rate for the model (Picard and Cook 1984). If the error rate is very high, a different classifier can be chosen, or a higher quality data set can be used to train the model. While machine learning is an efficient framework for using multiple inputs to make predictions, the predictions are only as good as the training data. As a result, if the training data is not very high quality, the predictions won't be either.

Lastly, the model is applied to the test set. For instance, when implementing a hidden Markov model, the forward-backward algorithm can be used to estimate the probability of class labels given a sequence (Majoros 2007). This can be useful when predicting TSSs if the class labels are $\{0, 1\}$, where 0 equals no TSS and 1 denotes a TSS. In addition, the Viterbi algorithm can predict the most probable path of labels (Majoros 2007). When applied to TSSs, this would elicit the most probable locations of transcription initiation throughout the genome. While TSS prediction is the focus of this work, because of its straightforward and rigorous probabilistic framework, classification has been applied to numerous problems throughout genetics, engineering, and computer science.

2.3.3 Markov Models (MM)

A Markov model is a probabilistic framework that is often used as a generative classifier in machine learning. Various types of Markov models exist. The simplest

among them is a Markov chain created from the linear ordering of states. Probabilities are assigned for each state $\mathbf{x}_i$ emitting an observation $\mathbf{y}_i$. A common application of a Markov chain is to calculate the probability of a sequence of DNA, given the emission probabilities of A,C,G, and T.

A second type of model typically seen in genomics is an interpolated Markov model (IMM) (Salzberg et al. 1998). An interpolated Markov model is used for cases in which higher order probabilities are based on data with a small sample size. Instead of discarding these probabilities, an interpolated Markov model uses a weighted average of probabilities of different orders, based on a function of their sample sizes. This is a form of smoothing, as it reduces sampling error by using lower-order probabilities (Majoros 2007).

A third type of model is a hidden Markov model (HMM). In an HMM, the sequence of states is not known, although, some probabilistic function of it is. In our machine learning example, the observable variables $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$ are known for the training set. If the observable variables $\mathbf{x}$ are labeled, the states $\mathbf{y}$ are known. If the observable variables $\mathbf{x}$ are not labeled, the states $\mathbf{y}$ are inferred from the input sequences during training. We maximize the probability of the variables $p(\mathbf{x}|\mathbf{y})$ conditional on the states $\mathbf{y} = (\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n)$ for the classifier $\mathbf{f}$. Then, we apply the classifier to the test set of variables $\mathbf{x}$ to determine the sequence of the states $\mathbf{y}$. In this way, we can classify individual nucleotides corresponding to specific sites of transcription initiation.

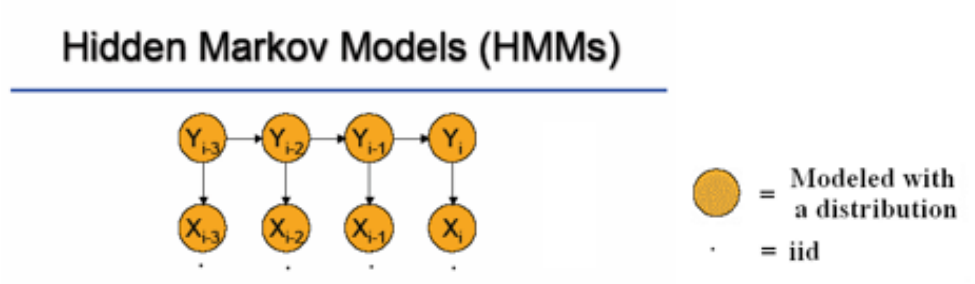


Figure 20: Hidden Markov Model

Nodes represent identically independently distributed (iid) variables $\mathbf{x} = (\mathbf{x}_{i-3},\mathbf{x}_{i-2},\mathbf{x}_{i-1},\mathbf{x}_i)$ and states $\mathbf{y} = (\mathbf{y}_{i-3},\mathbf{y}_{i-2},\mathbf{y}_{i-1},\mathbf{y}_i)$.

We can also model features of the genome longer than one nucleotide, such as introns or exons. As certain lengths may be more probable than others, to avoid length bias, a duration distribution can be incorporated into the model to denote the probability that the emission of a state will be of length *d* (Majoros 2007). A HMM with a duration distribution is called a generalized hmm (GHMM). GHMMs have played a central role in gene finding across species, in such programs as GENIE (Kulp et al. 1996) and GENSCAN (Burge and Karlin 1997). In *Drosophila*, GHMMs were used to provide the foundation of initial gene annotations in flybase (Wilson, Goodman, and Strelets 2008). A 5th order IMM having an overall GHMM structure was implemented in a program called McPromoter to predict promoters using the DNA sequence (Ohler et al. 2001). Most recently, GHMMs have proven successful in achieving a higher accuracy of

promoter prediction using modules of motifs in core promoters as states (see Figure 21) (Ohler 2006).

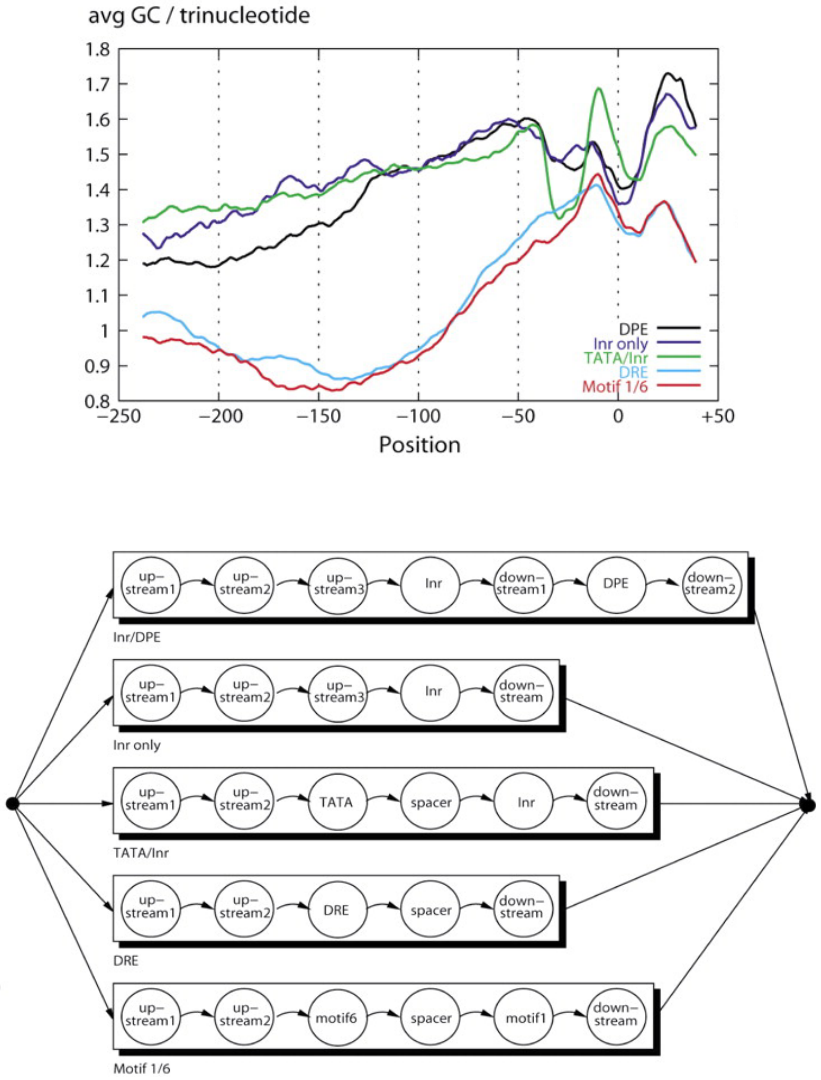


Figure 21: Promoter Prediction in *D.melanogaster* Using a GHMM

Combinations of core promoter motifs and sequence features were used as states to model promoter locations in a GHMM. Certain pairs, or modules of motifs, achieved the highest accuracy (right), and revealed two distinct profiles of GC

content (left). The DPE, INR, TATA/INR had elevated GC upstream of the TSS, while the DRE and Motif 1/6 showed a lower enrichment (Ohler 2006).

Markov models provide an effective statistical framework that can be applied to various domains. They are statistically elegant and easy to understand. As with all machine learning techniques, one must be careful when estimating parameters from the data because the predictions are only as good as the training data, and with a small amount of training data, other methods can outperform Markov models. Until recently, *Drosophila* 5′ TSS data was noisy and limited in availability due to the imprecision of the transcriptional machinery and experimental error. In addition, if one wants to combine different input data, the variable inputs of Markov models must be independently generated and probabilistically modeled. This creates challenges when trying to identify TSS locations because 5′ data, such as ESTs and tiling arrays, are not independent, and ESTs provide counts, rather than probabilities. Furthermore, it is computationally intensive to use Markov models for higher order dependencies, such as promoter motifs far upstream of the TSS.

2.3.4 Support Vector Machines (SVMs)

A Support Vector Machine (SVM) is a discriminative classifier in machine learning. If $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$ corresponds to feature variables, and $\mathbf{y} = (y_1,y_2,\ldots,y_n)$ corresponds to class labels, a SVM models the function f: $\mathbf{x}$ -> $\mathbf{y}$. A SVM is applied to

binary class labels, so for i = 1,…,n, $y_i = \{-1,1\}$, instead of a multi-dimensional vector. The SVM uses *Lagrangian multipliers* and a distance function called a kernel **k**(*,*) to compute the dot product of the variables in a high dimensional feature space (Majoros 2007). Figure 22 shows the dot products of the variables xi, xj mapped into the n and n+1 dimensions of the feature space (Schafer 2010). It is important to note that the n dimensions of the feature space in this specific illustration do not correspond to the n dimensions of the training data **x**,**y** generalized earlier.

Figure 22 shows that the dot products of the training data are plotted according to their class labels. Navy blue circles correspond to the dot products of variables with the class label -1, representing the inactive class, and light blue squares correspond to the dot products of variables with the class label +1, representing the active class (Schafer 2010). Regression is performed on the dot products to estimate a hyperplane that provides the highest amount of separation between the two sets of class labels. Support vectors perpendicular to the hyperplane are used to establish this distance, or margin, from the hyperplane to the closest dot products from both sets. Optimization is used to learn these parameters, and dot products within the margin, or on the opposing side of the hyperplane, are considered errors and used to measure how well the model fits the training data.

Then, the SVM can be used to classify the test data. Variables having dot products greater than the hyperplane are assigned one class label, while variables with

dot products less than the hyperplane are assigned the other. If a clean split between the variables cannot be made, slack variables are introduced that measure the degree of misclassification of the data. The goal of an SVM is to maximize the margin, minimize the error rate, and minimize the risk of overfitting (Burges 1998).

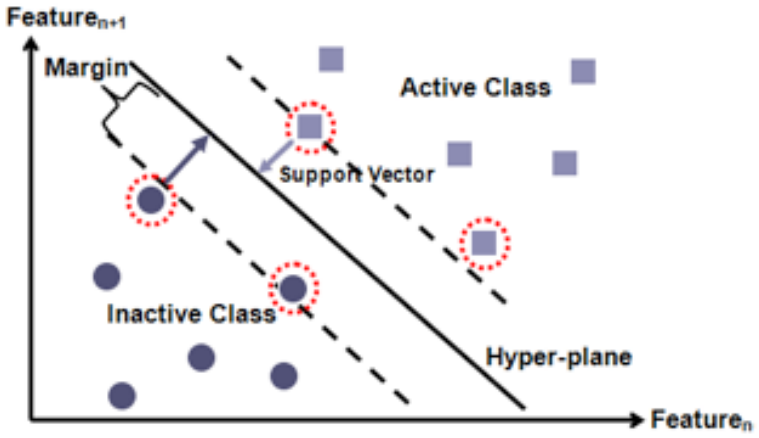


Figure 22: SVM Modeling the Separation of the Active and Inactive Class Labels by the Hyper-plane

After the parameters of the SVM are learned, variables in the test data with dot products greater than the hyper-plane are assigned the 'active class' label, while those less than the hyper-plane are assigned the 'inactive class' label (Schafer 2010).

SVMs have been used to model various genomic features, including individual exons and translational start codons (Majoros 2007). They have also been used in conjunction with HMMs to classify unannotated promoters in *S.cerevisiae* (Pavlidis et al. 2001). In human, transcription starts have been identified using the SVM based program called Accurate Recognition of Transcription Starts (ARTS) (Sonnenburg, Zien, and

Ratsch 2006). While a few studies have applied SVMs to promoter prediction in *D.melanogaster*, one used a 4-mer SVM based approach to predict polII bound promoters across 5 species, including the fruit fly (Anwar et al. 2008).

SVMs offer various advantages over alternative models. By choosing the maximum margin of the hyper-plane to fit the test data, overfitting is avoided. In addition, the model is sparse, and does not need to search for a local maximum. SVMs can easily map data into a high dimensional feature space and achieve higher accuracy than other methods. The greatest advantage of SVMs is that they locate a hyperplane without the computational burden of having to explicitly represent it. In spite of these benefits, SVMs can be slow to run, the weights on the features are non-transparent, and like all classifiers in machine learning, the accuracy of the predictions is dependent on the quality of the training data.

2.4 Computational Identification of Promoter Motifs

2.4.1 Position Weight Matrices (PWMs)

The computational techniques discussed thus far have focused on identifying TSSs directly from features of the genome. TSSs can also be identified indirectly by finding motifs in their promoter sequence that are essential in transcription initiation. Motifs were first discovered in the late 1970s by experimentally deleting regions of sequence upstream of the start codon and evaluating if the promoter was still functional

(Lifton et al. 1978). Since then, motif finding has grown to include various computational techniques (see sections 2.4.2, 2.4.3, 2.4.4) that can elicit strings of nucleotides represented more frequently in the sequence than expected by chance. These techniques have been successful in eliciting novel regulatory motifs in a variety of studies, including the discovery of the TATA and INR motifs in the fruit fly (Ohler et al. 2002). Motifs that are found are believed to have functional implications on the regulation of transcription, and can be used to identify genes with similar expression profiles, and to estimate the locations of TSSs.

Binding sites are typically represented as matrices called position-specific scoring matrices (PSSM), or position weight matrices (PWMs). Both nomenclatures signify strings of letters that symbolize the order and frequency of nucleotide occurrences. The strings may be of any length, although, they typically range from 5 to 15bp, the typical size of a transcription factor binding site. Each letter within the string corresponds to a vector $\mathbf{M} = \{M_{iA}, M_{iC}, M_{iG}, M_{iT}\}$ of observing all nucleotides $j = \{A,C,G,T\}$ at position i. The values M_{ij} often represent as log-likelihoods, $\log(p_{ij})$, or as log-odds scores, $\log(f_{ij}/B_j)$, reflecting the log odds of observing the frequency f_{ij} of a nucleotide above a background model B_j. The positions within a PWM are assumed to be independent of each other. Higher order dependencies can be created by pooling positions together in a higher order model called a weight array matrix (WAM) (Zhang and Marr 1993).

PWMs are often represented using weblogos, in which the overall heights of the nucleotides differ at each position, reflecting the variation in information content. Information content is measured in bits and can be thought of as the deviation of a PWM from a uniform distribution of nucleotides. Information content is derived from entropy, and can be calculated as (Schneider and Stephens 1990):

$$H(i) = - \sum f_{i,j} \log_2 f_{i,j}$$

One of the most popular programs used to create weblogos is the Berkeley website http://weblogo.berkeley.edu (Crooks et al. 2004).

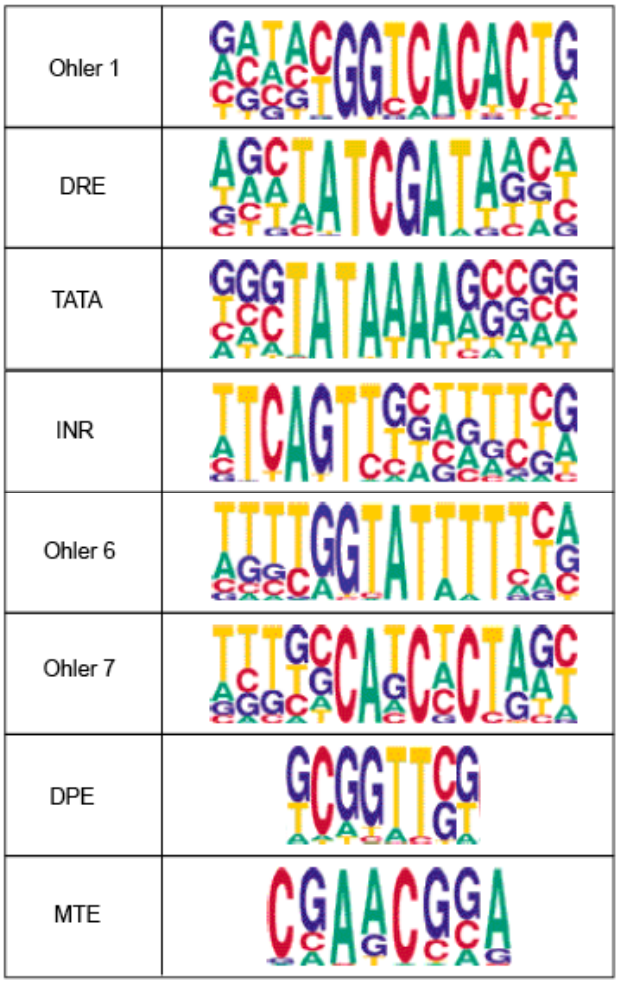


Figure 23: Pictogram of the PWMs of Core Promoter Motifs in *D.melanogaster* (Ohler et al. 2002)

The motifs are the Ohler 1, Downstream Replication Element (DRE), TATA box (TATA), Initiator (INR), Ohler 6, Ohler 7, Downstream Promoter Element (DPE), and the Motif Ten Element (MTE).

Illustrations of PWMs in which all of the nucleotides have been scaled to the same height are called pictograms. Figure 23 shows a pictogram for the PWMs of eight of the ten most overrepresented core promoter motifs in *Drosophila* (Ohler et al. 2002).

Because PWMs provide independent motif probabilities at each position, this degeneracy limits information about the most probable combinations of nucleotides across the string. For instance, if a PWM representation is A(T/G)(C/A), it is difficult to determine what 3-mer is most common: ATC, ATA, AGC, or AGA. These nucleotide variations may make a difference in the affinity and type of transcription factors binding to the DNA.

The presence of a motif in a sequence can be evaluated using a sequence comparison program, such as PATSER (Hertz and Stormo 1999). In PATSER, the sequence to be searched is divided up into overlapping subsequences of length L. Each L-mer is compared against the PWM, and given a score. Then, a p-value is calculated as the probability of observing a particular score or higher at that position in background sequences. If the p-value is smaller than the user specified cutoff, the location of the L-mer is returned as a match to the PWM. The process must be repeated for comparisons to additional PWMs. In this way, the complete set of motif locations in the sequence can be elicited.

2.4.2 Expectation Maximization (EM)

In the identification of motifs, there are two unknown variables $\theta = (\mathbf{M}, \boldsymbol{l})$, the nucleotides at each position in the PWM $\mathbf{M}$ and the motif locations $\boldsymbol{l} = (l_1, l_2, \ldots, l_d)$ in sequences $S_1, \ldots, S_k$. The number of motif locations d does not have to equal the number of

sequences k, although, often one motif per sequence is assumed, and $d = k$. One of the techniques employed to estimate these parameters is called Expectation Maximization (EM). The EM algorithm returns the most likely motif in the sequences using expected motif locations. The algorithm begins by selecting an initial guess for the PWM $\mathbf{M}^0$ and random locations $\boldsymbol{l}^0$ in the sequences. In the E step, the expectation of the log likelihood is determined using the initial parameters,

$$Z^0 = E(\log L^0) = E(\log (p(\text{motif} \mid \mathbf{M}^0, \mathbf{l}^0)).$$

In the M step, the log likelihood is maximized by resetting $(\mathbf{M}^0, \mathbf{l}^0)$ to their expected values $(\mathbf{M}^1, \mathbf{l}^1)$. This elicits a modified PWM $\mathbf{M}^1$, which is used to find the most probable second set of motif locations $\mathbf{l}^1$. Then, the expectation of the log likelihood is calculated again using the updated parameters:

$$Z^1 = E(\log L^1) = E(\log (p(\text{motif} \mid \mathbf{M}^1, \mathbf{l}^1)),$$

and the values for $(\mathbf{M}^1, \mathbf{l}^1)$ are reset as the expected values $(\mathbf{M}^2, \mathbf{l}^2)$. This process is performed for multiple iterations $t = 0,1,\ldots$, repeatedly choosing the model parameters $\boldsymbol{\theta}^{t+1}$ that maximize the expectation Z^t. The probability of a motif λ is found at each iteration and when λ converges, $\lambda^t - \lambda^{t-1} < \varepsilon$ for $\varepsilon > 0$, the most probable PWM and locations of the motif are estimated as $\boldsymbol{\theta}^* = (\mathbf{M}^*, \mathbf{l}^*)$.

MEME is the most well known program for implementing the EM algorithm in motif identification to date (Bailey and Elkan 1994). It has been in existence for 16 years, and has been incorporated into various suites of motif identification tools, such as the

Binding-site Estimation Suite of Tools (BEST) (Che et al. 2005). MEME has been employed in the discovery of numerous motifs, including the identification of the eight essential core promoter motifs in *D.melanogaster* listed in the previous section (Ohler et al. 2002).

The EM algorithm has provided a foundation for motif identification. Its direct interpretation and easy implementation have facilitated its use by scientists with only a minimal understanding of motif discovery. The EM algorithm has certain limitations that pose challenges to the accuracy and biological interpretability of its results. When employing the EM, an expected number of motif occurrences per sequence must be specified a priori. As these parameters are unknown, spurious motifs can be returned from inaccurate initial estimations. Models incorporating more than one occurrence per sequence are available, however, more complex statistics are needed for their computation, and this has led many studies to employ the unrealistic biological model of one motif per sequence. The biggest downfall of the EM algorithm is that it is similarly influenced by the choice of starting parameters $\theta^0 = (\mathbf{M}^0, \mathbf{l}^0)$, which does not guarantee that it will reach a global maximum during optimization.

When the EM algorithm is used to search for multiple motifs, motifs are returned successively one at a time, and their sequence is removed from sequential motif searches. This limits the simultaneous discovery of motif modules that may be cooperatively utilized, and does not prevent all motifs with highly similar position

weight matrices from being returned. On account of these caveats, the EM algorithm is often run more than once to properly adjust the starting parameters, and the results are compared to and verified by additional motif search techniques.

2.4.3 Gibbs Sampling

Gibbs sampling is a Bayesian approach to motif finding that was named after the statistical physicist J.W. Gibbs. Instead of providing a closed form solution, Gibbs sampling is used to approximate the posterior probability of the two unknown variables, the motif's PWM and locations $\theta = (\mathbf{M}, \boldsymbol{l})$, when only the conditional probabilities of θ given individual sequences can be sampled. Gibbs sampling was derived from three previous mathematical theorems. The first was Brook's Lemma of 1964 that showed

if $f(y_i \mid y_j, j \neq i)$ for $i = 1,\ldots,n$,

then $f(y_1,\ldots,y_n)$ is uniquely determined

(Gelfand and Vounatsou 2003).

The second was Bayes' Theorem

$$p(y) = p(y \mid x)\, p(x)$$ (Wassarman 2004).

And the third was the Metropolis-Hastings algorithm that showed

if x_i are drawn from $p(x)$,

then $E_{p(x)}[f(y \mid x_i)] = 1/n \sum f(y \mid x_i)$ (Wassarman 2004).

In Gibbs sampling, the most probable PWM and locations of the motif are estimated as $\boldsymbol{\theta}^* = (\mathbf{M}^*, \mathbf{l}^*)$ (Wassarman 2004). There are priors on the motif and its locations in the sequences, and you use the sequence to update the priors by evaluating the *maximum a posteriori* (*MAP*). Like the EM algorithm, Gibbs sampling begins by randomly selecting one motif location l^1 in the sequences $\mathbf{S} = S_1,\ldots,S_k$ and an initial guess for the PWM $\mathbf{M}^0$. Then, one sequence S_1 is removed from the set, and the log-odds ratio of observing the motif in sequences $S^1 = S_2,\ldots,S_k$ versus the background model is calculated as a first approximation of the PWM,

$$\mathbf{M}^1 = \sum \log (f_{ij} / B_j).$$

Then, a second location is sampled l^2, and the PWM is estimated a second time. This method of sampling is repeated multiple times, and the *MAP* of the model $\boldsymbol{\theta}^* = (\mathbf{M}^*, \mathbf{l}^*)$ is found by

$$p(\boldsymbol{\theta}^*, \mathbf{M}^*, \mathbf{l}^* \mid \mathbf{S}) = \frac{p(\mathbf{S} \mid \boldsymbol{\theta}, \mathbf{M}, \mathbf{l})\, p(\boldsymbol{\theta}, \mathbf{M}, \mathbf{l})}{p(\mathbf{S})}.$$

Sampling terminates when the *MAP* converges.

AlignACE is a commonly used implementation of Gibbs Sampling for motif identification (Hughes et al. 2000). In *Drosophila*, Gibbs sampling was used in the identification of the eight core promoter motifs (Ohler et al. 2002). Some disadvantages of Gibbs sampling are similar to that of the EM algorithm. The unknown motif size must

be specified a priori, and when searching for multiple motifs, the algorithm returns one at a time. One advantage of Gibbs sampling is that it is done from a parameterized distribution, so it is possible to achieve global convergence. Although, it can take a long time for the model to converge, often leading researchers to use sub-optimal solutions. The greatest advantage of Gibbs sampling is that it is more versatile, and easier to enhance. For applications in genomics, additional extensions of Gibbs sampling have been developed, such as nested sampling in the program NestedMICA (Down et al. 2007), and PhyloGibbs-MP for motif identification across lineages (Siddharthan 2008).

2.4.4 Position Overrepresentation

Alternative methods to the EM algorithm and Gibbs sampling have been applied to motif finding. One approach has been to search for motifs based on their position overrepresentation from a reference point. In this method, the user inputs the length n of the motif, and all possible combinations of nucleotides at each position are generated to create n-mers in an exhaustive fashion. Then, the nucleotide patterns are found in the sequences, and those with the highest enrichment from a reference point, above what would be expected by random chance, are returned.

Position overrepresentation models offer the advantage of incorporating a location parameter into motif searches, instead of solely relying upon the DNA sequence. The location parameter may elicit more biologically meaningful results, as

certain genomic features are arranged at regular intervals, such as pairs of transcription factors that have been shown to bind in coordination with each other, such as the TATA and INR in the fruit fly (FitzGerald et al. 2006; Ohler 2006). The importance of position overrepresentation for transcription factor binding at promoter local regions has begun to be explored in the human genome. One approach used position specific motifs to predict the set of narrow well-defined promoters (Megraw et al. 2009), and another used the position specific DNA sequence to accurately model precise TSS locations (Frith et al. 2008). Minimal applications of position specific models have been employed in *Drosophila*.

One outcome of position overrepresentation models is that the quality of the results directly reflects the accuracy of the reference point, which can be beneficial in promoter searches for well-aligned TSSs. Although, as TSS locations in *Drosophila* were poor to mediocre until recently, this has historically decreased the power of finding biologically functional motifs. Other caveats of these models include their long running time due to the exhaustive nature of the n-mer searches, and the input sequences are required to be the same length, which may be unrealistic for promoters of varying sizes that contain different combinations of motifs.

2.5 Experimental and Computational Methods Are the Yin and Yang of Science

In light of the advantages and disadvantages of the experimental and computational methods, the best results are achieved through the use of various cross-disciplinary techniques. The segregation of wet lab scientists from computational gurus remains a history of the past. The most insightful results have been achieved from the repeated cycle of producing experimental data to create mathematical models that are used to more accurately generate data to improve model performance. The hallmark of next generation computational biologists has been to merge the biological and computational sciences into one, akin to the yin and yang of Chinese philosophy. In this work, I aim to achieve this balance through the study of the spatiotemporal code of transcription initiation in *D.melanogaster*.

3. Identification, Motif Composition, and Conservation of Promoter Patterns

Elizabeth Rach conceived and performed all of the work in this chapter, except for the conservation analysis of core promoter motifs across the 12 *Drosophila* Genomes, which was contributed by Dr. Uwe Ohler. The work was published in *Genome Biology* in July 2009.

3.1 Introduction

Transcription is a crucial part of gene expression that involves complex interactions of *cis*-regulatory sequence elements and trans-factors. It is mediated in large part through the binding of transcription factors (TFs) to DNA sequence motifs. The majority of eukaryotic genes (protein-coding genes and many regulatory RNAs) are transcribed by RNA polymerase II (RNA pol II), an enzyme that contains various subunits and can exist in a holoenzyme complex with several basal TFs, including TFIIB and TFIIF (Latchman 2005). As RNA pol II does not have a direct affinity for the DNA, general TFs that bind to sequence motifs in the 100-bp region immediately surrounding the transcription start site (TSS), called the core promoter, guide it to the site of transcription initiation (Ohler and Frith 2005; Smale and Kadonaga 2003). The set of general TFs includes TFIID, which consists of the TATA-box binding protein (TBP) and 10 to 14 TBP-associated factors (TAFs), along with TFIIH, and others.

The availability of whole genomes and large-scale transcript data for different species has increasingly shown that the previously known canonical motifs were by far not frequent enough to constitute one general model of core promoter structure (Ohler et al. 2002; Suzuki et al. 2002). In particular, recent high throughput sequencing efforts based on 5′ capping protocols have now generated capped transcripts for human and mouse on a high throughput scale (Carninci et al. 2005; Kimura et al. 2006; Valen et al. 2009). These "5′-capped" or "cap-trapped" transcripts have helped to identify genomic TSS locations for thousands of genes, in particular for human, mouse and yeast (Carninci et al. 1997; Schmid et al. 2006; Zhang and Dietrich Mapping of transcription start sites in saccharomyces cerevisiae using 5' SAGE 2005).

This approach revealed that transcription is often initiated across widespread genomic locations, making it non-trivial to define initiation sites (Carninci et al. 2005; Carninci et al. 1997; Kawaji et al. 2006; Kimura et al. 2006; Schmid et al. 2006; Zhang and Dietrich Mapping of transcription start sites in saccharomyces cerevisiae using 5' SAGE 2005). Two general initiation patterns have been characterized in mammalian core promoters. The first contains those with tags mapping to a "single dominant peak," whose promoters have strong over-representations of canonical motifs, such as the TATA box, GC box, CCAAT motif, and comparatively low frequencies of CpG islands. Gene ontology (GO) analyses have shown that single dominant peaks are associated to developmental regulation and specialized differentiation processes (Carninci et al. 2006).

The second type of initiation patterns is "broad regions" whose promoters have TATA poor profiles and are enriched in CpG islands. Broad regions are associated to more ubiquitously expressed transcripts with housekeeping functions, such as RNA processing and the ubiquitin cycle (Carninci et al. 2006). The large scale of available data allows for detailed analyses; for instance, one study explored the importance of precise spacing between the TATA and the TSS (Ponjavic et al. 2006).

Until recently, data comparable in scope to the CAGE sets for mouse and human has not been available for *Drosophila* genomes (Clark et al. 2007; Stark et al. 2007), but a large number of ESTs have been sequenced in *D. melanogaster* using 5′ capping technology (Celniker et al. 2002). Using these, several computational efforts have focused on the locations and frequencies of sequence motifs found in core promoters. The TATA box (TATA), Initiator (INR), Downstream Promoter Element (DPE), and Motif Ten Element (MTE) have been identified with distinct spacing requirements relative to the TSS (Juven-Gershon et al. 2008). Each of these motifs has been found at a comparatively low frequency, but several analyses have identified common additional motifs enriched in core promoters (FitzGerald et al. 2006; Ohler et al. 2002). A different analysis showed that specific motif combinations, or modules, frequently occur in core promoters (Ohler 2006). These modules are hallmarks of distinct core promoter types, and have been shown in a study of genes associated with highly conserved non-coding element to characterize three main functional classes of genes in *D. melanogaster*:

developmental regulation, housekeeping, and tissue specific differentiation (Engstrom et al. 2007) . Such functional classes have also been associated to different modes of RNA pol II occupancy (Zeitlinger et al. 2007).

The core promoter elements and modules also offer deeper insight into the combinatorial utilization of the sequence architecture. For example, the enhancer for the *yellow* gene has been shown to interact with a promoter in cis and a promoter in trans (Lee and Wu 2006). With respect to higher levels of organization, genomic analyses are increasingly complemented by the elucidation of epigenetic patterns, such as the positioning of nucleosomes and the presence of certain histone marks (Mavrich et al. 2008; Mito, Henikoff, and Henikoff 2005). Previous analyses used polytene chromosome staining and ChIP-on-chip to show the existence of two distinct transcriptional programs in *D.melanogaster*: the TATA-box-binding protein – related factor 2 (TRF2) regulating TATA-less transcription, including the genes encoding linker histone H1, and the TATA-box-binding protein (TBP), including transcription of promoters of the core histones H2A/B, and H3/H4 (Isogai et al. 2007). The degree to which the core promoter motifs/modules and epigenetic features are correlated with the patterns of transcription initiation has not yet been explored in *D.melanogaster*.

In this chapter, we use available large-scale data to provide an extensive, high-quality mapping of alternative TSSs across the fruit fly genome. We show that individual core promoter elements and their corresponding modules are associated to

the peaked and broad patterns of transcription initiation that characterize them. Lastly, we confirm that motif matches are highly conserved in the peaked promoters of TSSs, but show considerable variation in the broad promoters of TSS cluster groups.

3.2 Materials and Methods

3.2.1 EST Filtering and Clustering

We used EST alignments from *Drosophila* Release 4.3 to identify TSSs, which enabled us to directly map our results to other available data sources. We filtered the ESTs in a four-step process by first eliminating ESTs that did not cover an intron splice junction. This reassured us that the remaining ESTs were produced from mature transcripts. Second, we removed ESTs having aligned fragments longer than 1,500nt, or a distance greater than 100kb between any two fragments. This was done to exclude dubious ESTs that may incorrectly map to the genome. The parameter range of 50-100kb corresponded to an upper bound of the genomic span of fly genes and was previously used as a natural cutoff for the determination of promoter co-regulation (Manak et al. 2006) . Third, we took out ESTs that aligned to multiple regions to ensure our set contained unambiguous locations. Fourth, we deleted ESTs with the most 5′ location mapping to within 2bp of the start of a downstream exon or transposon, as annotated in Release 4.3. This served to eliminate incomplete ESTs, and those utilized by transposons.

The 157,093 ESTs that remained were deemed highly confident in mapping to the most 5′ ends of coding transcripts.

We implemented a hierarchical clustering strategy to define individual TSSs (Figure 25). We first parsed the ESTs by associating each of the 157,093 filtered ESTs to corresponding genes and dividing all of the ESTs for each gene into broad windows. Adjacent ESTs that were less than 100bp apart were assigned to the same window, while adjacent ESTs greater than 100bp apart were assigned to different windows. The window size of 100 nt is a rule-of-thumb standard which has also been employed by EPD to specify broad regions of transcription initiation (Schmid et al. 2006). Moreover, the known sequence features directly involved in transcription initiation are all located within +/-50 nt from the TSS, and the core promoter region of each TSS is generally defined to be ~100bp in size. The genomic position of the 5′ end of each EST alignment is referred to as the EST location.

We next computed the standard deviation of EST locations, and iteratively divided windows into smaller clusters until each had a standard deviation of less than 10. We will refer to all of the clusters and sub-clusters having a standard deviation less than 10 with the term (sub-) cluster. This was done to discriminate regions of high localized EST frequency from broad regions with low EST frequency. It also served to separate singleton EST outliers into separate (sub-) clusters. The choice of 10 as standard

deviation parameter corresponds to a variance of 100bp and thus the size of a core promoter, as defined above.

3.2.2 TSS Identification From EST Clusters

We identified TSSs from the (sub-) clusters using four criteria. First, we found the location with the highest frequency of ESTs in each (sub-) cluster, and removed (sub-) clusters with a maximum frequency at a single site less than 2. This criterion selected only those (sub-) clusters with consistently and reproducibly utilized TSSs. If 2 or more sites were tied for having the highest frequency of ESTs, the upstream site was chosen.

Second, to ensure that predicted locations coincided with the beginning of full-length transcripts, we selected sites that had to either be supported by at least three ESTs from a 5' capped library sequenced by RIKEN (Carninci et al. 2005), or two RIKEN ESTs and a third EST within 5bp from any non-RIKEN, non-capped library. For EST clusters without RIKEN ESTs, sites had to either be supported by three ESTs within 5nt of the 5' end of the cluster, or have at least half of the ESTs within a (sub-) cluster falling within 5nt of each other.

Third, if a cluster contained several TSSs identified for more than one (sub-) cluster, we placed a new window starting at one TSS and ending at the second TSS. If the standard deviation of this new window was less than the cutoff of 10, we kept the

site with the higher frequency of ESTs as the TSS and removed the second location from the dataset. If the standard deviation of the new window was greater than 10, we kept both locations as TSS candidates. This eliminated closely spaced TSSs from adjacent (sub-) clusters.

Fourth, we required sites to be upstream of a start codon annotated for the gene in Release 4.3. Because ESTs do not span the entire length of a transcript, we generally do not know what downstream isoforms correspond to the TSSs. For this reason, we conservatively required TSSs to be upstream of the most downstream start codon. If any of these criteria were not satisfied, we declared the (sub-) cluster to not have any conclusive TSSs and removed it from further analysis.

3.2.3 Core Promoter Motif and Conservation Analysis

We applied the program PATSER (Hertz and Stormo 1999) to the plus strand of the core promoter region [-60,+40] bp immediately surrounding the identified TSSs and the most 5′ sites in Flybase, to look for hits to previously published position weight matrices above a threshold. For broad TSS cluster groups, promoter sequence [-60] bp of the most upstream TSS to [+40] bp of the most downstream TSS in the cluster group was extracted. To assess the strength of enrichment, we extracted 100bp sets of sequences

surrounding three randomly selected intergenic sets of sites, and repeated motif searches on these sets.

We used relative frequency matrices for eight core promoter motifs reported by Ohler et al (Ohler et al. 2002) and that were confirmed by analyses of other groups, e.g. Fitzgerald (FitzGerald et al. 2006). We estimated set-specific mononucleotide backgrounds to account for varying AT content in the promoter sequences we analyzed (our TSS set; Flybase TSSs; and the random intergenic set). Score thresholds were individually chosen for each position weight matrix, always corresponding to a P-value of 10^{-3} for the expected false positive hit per nucleotide. As seen in Figure 31, motif matches in random intergenic regions agreed very well with the expected false positive rate. Motif matrices were taken from Ohler *et al.* (Ohler et al. 2002), with one modification. The DPE as reported in that study is a composite of the closely spaced MTE and DPE elements (this can clearly be seen when comparing Motif 9 (DPE) and Motif 10 (MTE) with previous DPE consensus motifs), which is likely a side effect of the MEME motif-finding strategy employed in that study. To avoid confounding results by overlapping matches, we shortened both DPE and MTE to 8nt non-overlapping motifs. All frequency matrices and background models are part of Appendix B.

Preferred motif positions were defined differently for location-specific and non-location specific core motifs: For TATA, INR, DPE and MTE, we used the 10nt window with the highest number of motif matches in our *D. melanogaster* TSS set (-38 to -29 for

the TATA box starting position, -4 to +6 for the INR motif, +14 to +23 for the MTE, and +21 to +30 for the DPE). These windows overlapped the most enriched motif locations as identified in the Flybase-defined promoter analysis of Fitzgerald et al. (FitzGerald et al. 2006). For the other four motifs, we used the 20nt windows as defined in that study (Ohler 1: -20 to -1; DRE: -60 to -41; Ohler 6: -60 to -41; and Ohler 7: +1 to +20). Note that we restricted motif matches to the preferred windows in some but not all analyses; in particular, preferred windows are somewhat less meaningful when dealing with broad cluster groups that do not exhibit a single initiation site.

For the conservation analysis, we first obtained orthologous regions across the other 11 species (Clark et al. 2007) using alignments computed by Multi-LAGAN (Brudno et al. 2003). Then, we selected promoters of TSSs having alignments in all 12 species, which led to a reduced set of 4,243 TSSs, with 2,075 genes with one TSS and 1,100 genes with more than one. As described above, we scanned orthologous regions in each species for motif hits above the threshold. For the location-specific motifs (TATA, INR, DPE, MTE), we identified matches in the *D. melanogaster* sequences within the 10-nucleotide preferred windows as defined above; for the other four motifs, we used the most-enriched 20-nucleotide windows (FitzGerald et al. 2006). Then, we assessed whether motif matches in *D. melanogaster* were located at corresponding positions in any of the other 11 genomes. Following the example of (Moses et al. 2006), we allowed for ±5 nucleotides to account for possible small errors in the local alignments at the site of a

motif match. In this way, we assessed whether a presumably functional motif, defined by the experimentally deduced location of the TSS and the occurrence of a motif match in the preferred position, was still detected in a second species, or potentially lost.

3.3 Results

3.3.1 EST Clustering Identifies a High Quality Set of Alternative Transcription Start Sites

Previous studies on *Drosophila* promoters have often been based on the analysis of upstream sequences extracted from a genomic resource such as Flybase (Wilson, Goodman, and Strelets 2008), using the most 5′ location of a gene as the site of transcription initiation. However, using a resource in this way invariably leads to inconsistent assignment of TSS locations; for instance, many Flybase transcript annotations begin with a start codon, indicating that no transcript evidence is available and making the annotation incomplete on the 5′ end. Filtering out such simple cases does not mean that the remaining transcripts are automatically 5′ complete. While the accuracy of TSS annotations have considerably improved with increasing available data (Wakaguri et al. 2008), the use of high throughput 5′ capping methodologies to identify TSSs has also revealed dispersed patterns of transcription initiation in mammalian genomes (Carninci et al. 2005; Kimura et al. 2006). These patterns have challenged the validity of choosing the most 5′ observed location as being the consistently utilized site.

Thus, we are not confident in the reliability and quality of TSS data extracted from general-purpose genomic annotations because we cannot be sure (1) which of the annotated 5'ends reflects a complete transcript, and (2) which ones accurately capture a true and consistently used transcription start site. Other previous analyses in *D. melanogaster* were based on high quality TSSs, but were smaller in size and depth. For instance, our previous core promoter study covered 1,941 TSSs, but did not include alternative start sites (Ohler et al. 2002). The Eukaryotic Promoter Database (EPD) incorporated highly confident TSSs identified from the curation of ESTs and is of a similar magnitude to our previous study (Schmid et al. 2004). Here, we continue the tradition of using ESTs for TSS identification, but with the goal of identifying all of the consistently utilized and precisely defined TSSs, rather than the most 5' ones.

To minimize experimental error and clearly distinguish true TSSs from background noise, it is essential to filter available 5' transcript data. To accomplish this, we started from the large dataset of *D. melanogaster* ESTs in the Berkeley *Drosophila* Genome Collection (BDGC) (Stapleton et al. 2002). Libraries were deposited by various research teams and cover a broad spectrum of different conditions (Figure 24). The Oliver Lab generated ESTs from the testes library BS, which is of interest because germline tissues exhibit a sex specific transcriptional program (Hiller et al. 2004; Metcalf and Wassarman 2007). Exelixis generated two immune-response related libraries, EN and EC, from LPS induced mbn2 cells, and fat bodies of third instar larva challenged

with gram +/- bacteria, respectively. There were 11 smaller libraries of ESTs grouped together into the library "OTHERS" and assigned the default condition diverse. ESTs from the BQ and EK libraries were also labeled with the diverse condition because they were derived from unknown and multiple sources (*Drosophila* embryos, imaginal discs, and adult heads). Some library conditions corresponded to developmental stages of the fruit fly life cycle (embryo and the larva/pupa). Schneider cells were captured by immortalized cell lines obtained from late embryonic stages (20-24 hrs). The remaining five specific conditions corresponded to body parts of the adult fruit fly. More information about the EST libraries can be found at the BDGC (Stapleton et al. 2002).

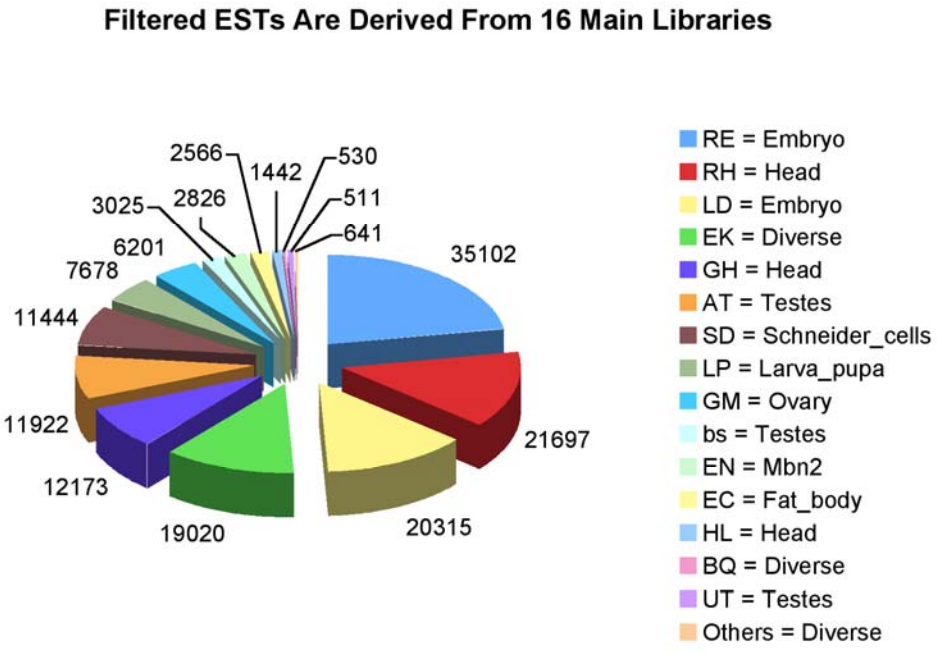


Figure 24: Sources of EST Data

631,239 EST alignments for 318,483 ESTs from the BDGC were taken from Release 4.3 of the fly genome annotation. The ESTs were filtered to a unique set of 157,093 alignments.

A significant fraction of ESTs were obtained with a protocol designed at the RIKEN institute to capture capped full-length transcripts (Carninci et al. 1997), similar to the more recent and larger mammalian efforts. This subset is therefore expected to map to the exact starting locations of known transcripts. While the amount of available ESTs is not large enough to completely saturate the transcriptome, it had until recently been the largest amount of transcript data for *Drosophila*. We mapped the BDGC ESTs derived from 15 different libraries to eight distinct conditions: embryo, larva/pupa, head, ovary, testes, Schneider cells, mbn2 hemocytic cells, and fat body. A broad adult stage can be accounted for by combining the promoter associations of the head, ovary, testes, mbn2 hemocytic cell, and fat body. Additional libraries from more than one body part or time period, an unknown source, or additional conditions than examined here, were assigned to one default condition called "diverse". By using independently generated cDNA libraries, we expect to reduce potential experimental biases from any one library in mapping TSS locations due to incomplete reverse transcription. This list of EST-library derived conditions is certainly limited, but it enables an initial analysis of promoter utilization in different life stages and differentiated tissues.

We started from a set of 631,239 EST alignments for 318,483 ESTs, which were part of Release 4.3 of the *D.melanogaster* genome. We filtered this initial set to a reduced

set of 157,093 unique EST alignments with high confidence of mapping to the 5′ ends of transcripts (see Materials and Methods). These unique EST alignments map across the *Drosophila* chromosomes and were derived from libraries of different sizes and conditions (Figure 24). The libraries providing the most ESTs were the RIKEN Embryo, RE with 35,102 ESTs, and RIKEN Head, RH with 21,697 ESTs. The remaining 100,294 ESTs were collected from non-cap trapping libraries. On account of the large size of the RIKEN libraries, the embryo and head conditions contained the largest number of ESTs, 55,417 and 35,312, respectively. ESTs mapping to the diverse condition, and those from the testes were next in size, followed by the Schneider cells, larva/pupa, and ovary. The mbn2 hemocytic cells and fat body conditions had the smallest numbers of ESTs.

3.3.2 Alternative TSSs Are a Widespread Phenomenon in the Fly Genome

To obtain a set of the most consistently utilized and precisely defined TSSs, rather than the most 5′, we implemented a hierarchical clustering strategy to define individual TSSs, summarized in Figure 25 (see Materials and Methods). We first associated each of the 157,093 filtered ESTs to corresponding genes, and then analyzed the distribution of ESTs for disjoint subsets, denoted "(sub-)clusters". We selected one or more TSSs from these (sub-)clusters for each gene using additional criteria (see Materials and Methods). All (sub-)clusters with less than 3 ESTs were removed from the analysis, and the individual TSS locations were required to be supported by at least two ESTs. By

designating TSS positions at the location of the highest EST frequencies within a clearly delineated cluster, instead of at every mapped location (Carninci et al. 2005) or the most 5′ one (Zhang and Dietrich Identification and characterization of upstream open reading frames (uorf) in the 5' untranslated regions (utr) of genes in saccharomyces cerevisiae 2005), we were able to gain new insights on the architecture of core promoters and their associations to conditions. The two most sensitive parameters in the clustering algorithm were the standard deviation and minimum frequency. We selected informed values based on previous analyses of *Drosophila* core promoters, however, increasing or decreasing these values changes the number of TSSs identified. Given the amount of available data and correspondingly chosen clustering parameters, all TSS positions are separated by at least 20bp, with the consequence that motif assignments, which were restricted to small, preferred windows in the core promoters relative to the TSSs, could be made to individual sites. In mammalian studies on large-scale 5′capped transcript datasets, initiation sites were observed at many closely spaced locations and called at single-nucleotide resolution (Ponjavic et al. 2006). While some of the initiation frequencies at this resolution have been shown to be condition-specific, broader TSS initiation patterns may potentially be a result of some degree of sloppiness in the transcriptional machinery, and functional consequences of such differences on transcription are of yet unclear. With the data used in this study, we could not resolve

whether start sites in flies are overall defined in tighter patterns, even if they are closely spaced to each other, or if additional data would lead to broader patterns.

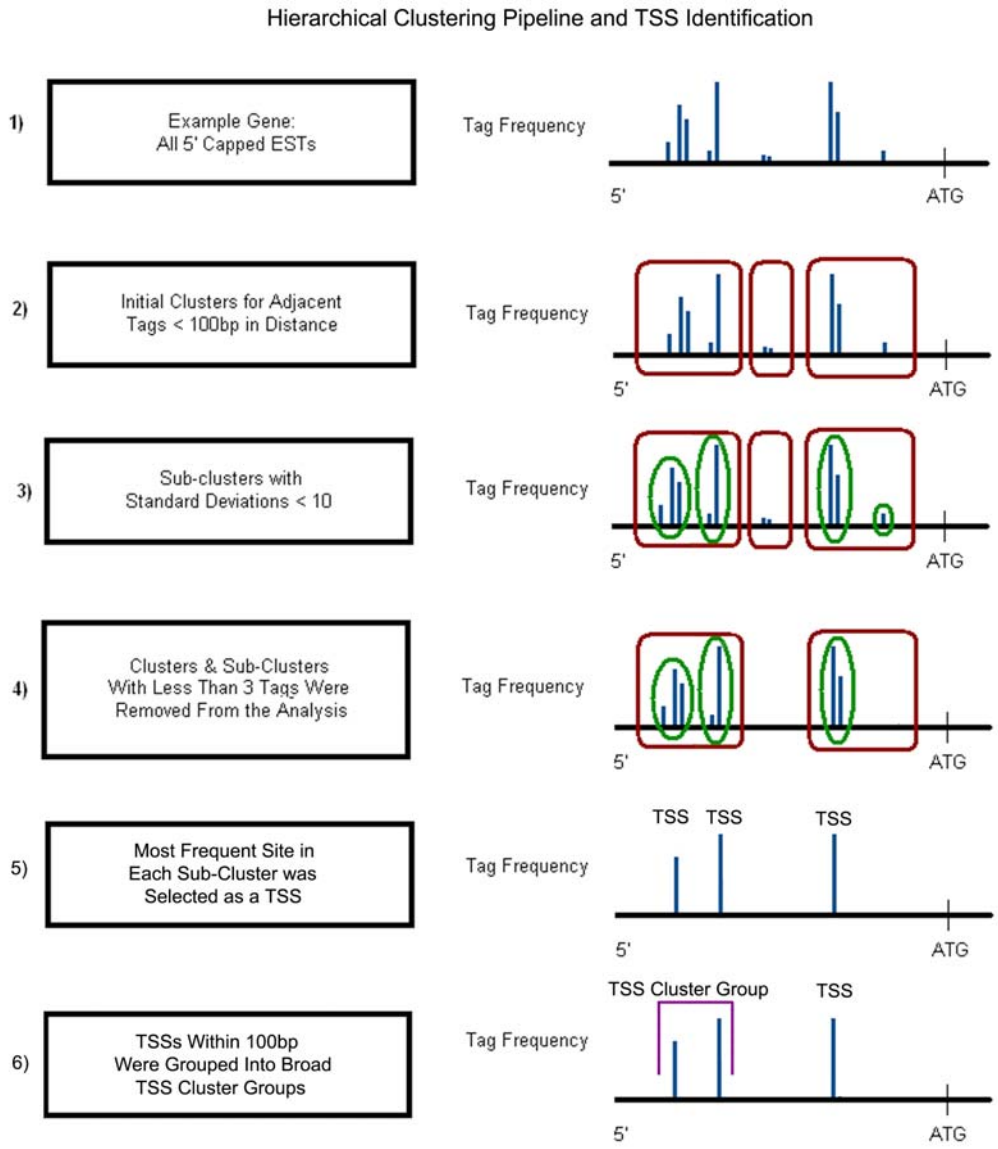


Figure 25: Hierarchical Clustering Algorithm and TSS Identification

ESTs were hierarchically clustered in 4 main steps. (1) ESTs were mapped to the 5′ ends of genes. (2) Large initial clusters were formed from grouping adjacent ESTs together that were less than 100bp apart. (3) Clusters were broken into

smaller (sub-) clusters that each had a standard deviation of less than 10. (4) (Sub-) clusters with less than 3 ESTs were removed. Then, (5) the most highly utilized location per (sub-) cluster was selected as the TSS and (6) TSSs within 100bp were grouped into broad TSS cluster groups.

We identified 5,665 TSSs for 3,990 genes (Appendix A), nearly three times the number of TSSs and twice as many genes as in our earlier study (Ohler et al. 2002). More than half of the filtered ESTs were removed in hierarchical clustering and TSS selection. The largest decrease in the number of ESTs during TSS selection was observed for the diverse category. This indicates that data from more variable sources show less consistent TSS locations compared to RIKEN cap-trapped data. TSS locations with overlapping core promoter sequences, i.e. less than 100bp from each other, were grouped into non-overlapping TSS cluster groups spanning longer promoter regions. Below, the TSSs in TSS cluster groups are analyzed on two levels: as individual initiation sites, and together in TSS cluster groups.

There were 2,765 genes (69%) with one TSS, and 1,225 genes (31%) with more than one TSS. Of the 1,225 genes with more than one TSS, 685 genes (56%) had one TSS cluster group, and 540 genes (44%) had more than one TSS separated by more than 100bp. Genes with alternative TSSs and alternative TSS cluster groups were distributed across the chromosomes 2L, 2R, 3L, 3R, and X (Figure 26). There may be additional alternative initiation sites upstream or downstream of those listed here that were not considered due to a lack of EST support.

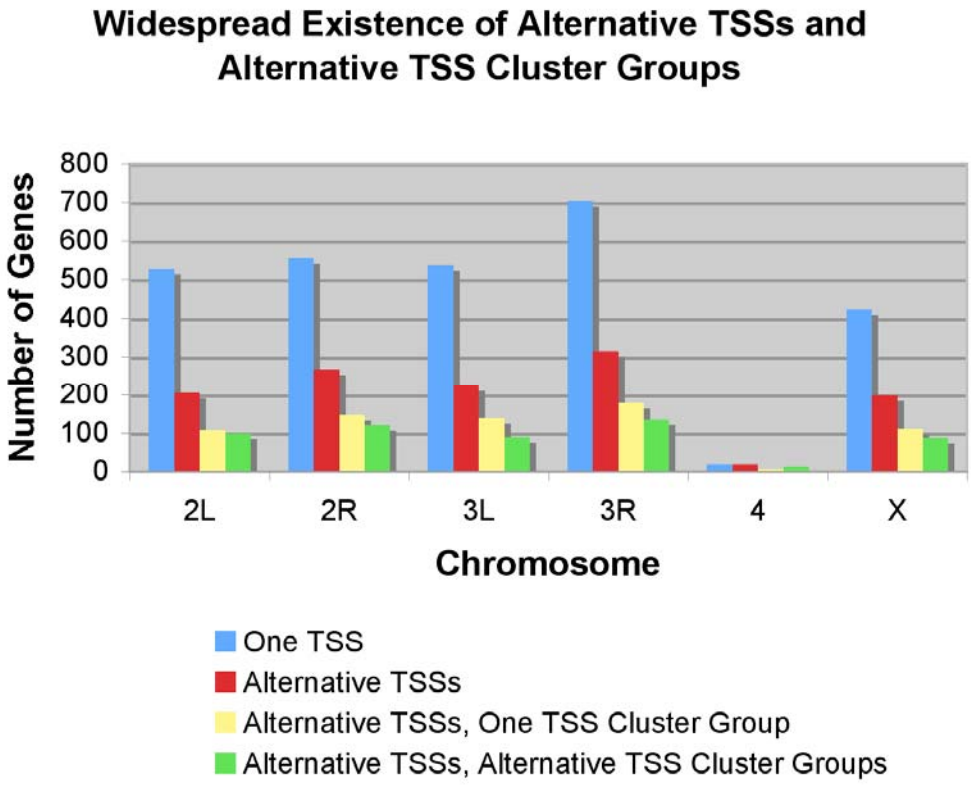


Figure 26: Alternative TSSs and Alternative TSS Cluster Groups Are Widely Distributed Across the Genome

For each chromosome, the number of genes with one TSS (blue) and more than one (i.e. alternative) TSS (red) were counted. Genes having alternative TSSs were divided into 2 groups: those having one TSS cluster group (yellow) and those having more than one TSS cluster group (green). With the exception of chromosome 4, the overall fraction of genes with alternative TSSs ranged from 28-32%, and the fraction of genes with alternative TSS cluster groups was 12-14%. Chromosome 4 is much smaller in size than the other *Drosophila* chromosomes, and had an elevated percentage of genes with alternative TSSs (19 out of 38, 50%) and alternative TSS cluster groups (34%), possibly due to the small sample size.

The mean genomic distance from TSSs to the most upstream start codon annotated in release 4.3 was 1,353bp, with a median of 264 bp. This is 91bp smaller than

our previous estimate of 1,444bp between TSS and start codon using Chromosome 2R (Ohler et al. 2002). This difference is likely due to the earlier strategy of Ohler *et al.* using the most 5′ ESTs to define sites of transcription initiation, rather than our use of the most highly utilized locations as TSSs. For genes with a consistent downstream start codon annotation, 141 TSSs were more than 10,000bp upstream of the closest start codon. This observation of large distances between TSSs and their corresponding start codons agrees with high frequencies of large distances between TSSs and start codons found in *D. melanogaster* using tiling arrays (Manak et al. 2006). Due to the clustering criteria, the minimal distance between two alternative TSSs was 20bp, with the most common distance ranging from 25-35bp. This is different from the more high-resolution definition of alternative TSSs that was employed in studies using high-throughput 5′ cap trapping data (Ponjavic et al. 2006). As a result, canonical core promoter sequence elements that occur at precise distances from the TSS such as the Inr, TATA box or DPE, can be clearly assigned to individual promoters.

The maximum number of individual TSSs identified per gene was seven for the genes CG33113 (*Rtnl1),* CG14039 (*quick-to-court)*, and CG11525 (*CycG)*. Flybase listed three fewer alternative TSSs for *quick-to-court,* and four fewer for *CycG* in Release 5.11(Wilson, Goodman, and Strelets 2008). Seven transcript isoforms for *Rtnl1* and *quick-to-court,* and three transcript isoforms for *CycG* are annotated for these genes. Whereas some of the TSSs of *CycG* and *quick-to-court* are close to each other and combined in

cluster groups, all of the TSSs of *Rtnl1* are well-separated peaked TSSs. Due to the stringent selection criteria we employed in the clustering strategy, genes with more than seven promoters may exist, but we found the most common range of alternative TSSs to be much lower.

Due to the definition of the TSS cluster groups, the minimal distance between TSSs in alternative TSS cluster groups is 101bp, and the most common intra cluster distance ranges from 101-199bp. There were 55 TSS cluster groups separated by more than 10kb. It is estimated that noncoding 5′ and 3′ DNA each comprise approximately 2kb of intergenic sequence, and that intergenic distances increase with regulatory complexity (Nelson, Hersh, and Carroll 2004). Genes performing house-keeping functions, such as ribosomal constituents and general TFs, are commonly spaced in 4-5kb segments of DNA. Genes with more complex roles, such as in embryonic development and/or pattern specification, take up 17-25kb of DNA on average. This suggests that some of the alternative TSSs/cluster groups separated by large distances may experience more complex transcriptional regulation.

3.3.3 Quality Assessment of Identified TSSs

3.3.3.1 Identified TSS Locations Correspond to Sites in Other Genomic Data Sources

We evaluated the quality of our set of alternative TSSs by comparing initiation locations to known sources. The first one we considered was the Eukaryotic Promoter

Database EPD (Schmid et al. 2006), which has long been the reference source for computational sequence analyses and modeling the proximal core regions of promoters. The TSSs listed in EPD are result from using the oligo-capped EST subset from the BDGC, a strategy parallel to ours. However, TSSs in EPD were identified with the program MADAP (Schmid et al. 2007), which requires a user-specified initial range of the number of TSS clusters, which is typically not known *a priori*. In contrast, the hierarchical clustering strategy that we used heuristically identifies TSSs using parameters based on genomic properties of core promoters, without any initial knowledge of the number of TSSs. Also, MADAP uses Gaussians to model EST distributions, regardless of fit, which may lead to calls of TSSs at the mean location instead of the most frequent one. In particular, isolated ESTs far away from a cluster have a large impact on the selection of the mixture model and may bias TSS locations, and may contribute to the differences observed between EPD and our set.

The promoters of TSSs in EPD are divided into three groups: those surrounding single transcription initiation sites, those around initiation regions, and those encompassing multiple initiation sites (Schmid et al. 2006). These categories correspond to peaked promoters, broad promoters, and alternative promoters in our set, and to single dominant peaked (SP), broad regions (BR), and more than one SP or BR in the vertebrate set (Carninci et al. 2006). EPD contains one-third the number of sites in our set, 1,926 TSSs, of which very few are alternative sites. To compare the locations of the

identified TSSs to those in the Eukaryotic Promoter Database EPD (Schmid et al. 2007), we downloaded 201 bp surrounding each TSS for 1,922 *D. melanogaster* promoters and aligned them to Release 4 with BLAST (Altschul et al. 1990). This was done to map entries to a common reference point because EPD does not use Flybase IDs for TSS identification. We removed 82 sequences with non-unique matches or less than 75% sequence similarity from the analysis, and compared the locations of the remaining 1,840 EPD entries to the closest TSSs in our set.

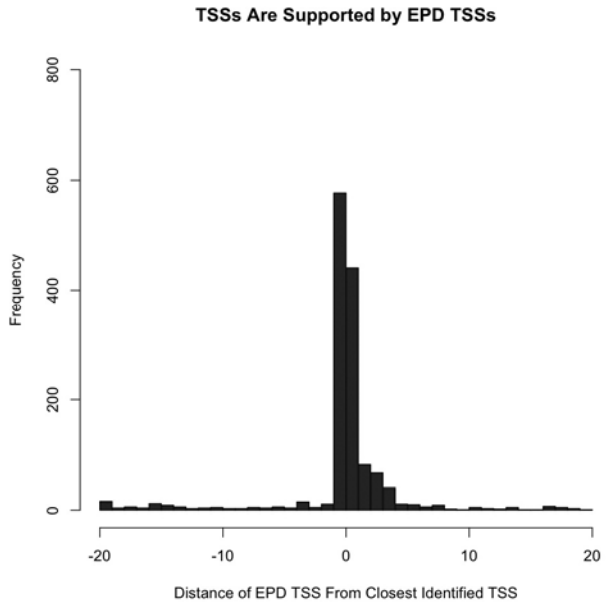


Figure 27: EPD Location Differences

Each of the 1,840 EPD TSSs was compared to the set of identified TSSs that were on the same chromosome. The difference in location of the closest identified TSS was taken from each EPD TSS, with the identified TSS as reference position, 0. Differences ranged from 0 to greater than 1,000 bp. The plot covers a region of +/- 20 nt, which covers 76% (1,404) of EPD start sites.

The distribution of distances between the locations of TSSs in our set and those in EPD showed a high percentage of overlap (Figure 27), with the mode difference being 0bp. The mean location of EPD TSSs was ~35bp downstream of our identified TSSs. There were 577 EPD TSSs (31%) that mapped exactly to the identified TSS locations and 1,404 EPD TSSs (76%) within +/- 20bp. The promoters of the set of 1,404 TSSs were distributed across all three EPD classifications in the following way: 35% were single initiation sites, 26% were multiple initiation sites, and 39% were initiation regions. There were 136 EPD TSSs that were located within 21bp to 1,000bp from the closest TSS, and 300 EPD TSSs were farther than 1,000bp away. Such large distances are likely the result from comparing the closest TSS locations between our set and EPD, regardless of different gene associations. Instead of disagreements in calling the location of a TSS, this fraction rather indicates the number of distinct TSSs that are included by either us or EPD alone, but not both.

The second source to which we compared our set of TSSs was Flybase. Flybase is the most complete database of *Drosophila* transcript information to date. Rather than TSS annotations *per se*, Flybase contains the 5′ ends of genes or transcripts which have been annotated using a variety of experimental and computational methodologies, followed by manual curation (Wilson, Goodman, and Strelets 2008). To provide a balanced comparison of Flybase TSSs to identified TSSs, the 5′UTR files of the same release (4.3) were chosen for analysis. Flybase TSSs were extracted for each chromosome, and only

those upstream of the gene's most downstream start codon, but not mapping to a start codon, were selected for comparison. These criteria were in agreement with the standards used to identify the hierarchically clustered TSSs, and ensured the exclusion of obviously suspicious or incomplete transcripts. A set of 18,767 TSSs for 9,655 genes from Flybase remained for comparison, over three times the number of TSS candidates in our set. Because we mapped TSSs to Flybase gene IDs, we were able to compare 5,610 TSSs for 3,945 genes in our set, ~99% of the TSSs identified by hierarchical clustering, to comparable sites of the same genes in Flybase.

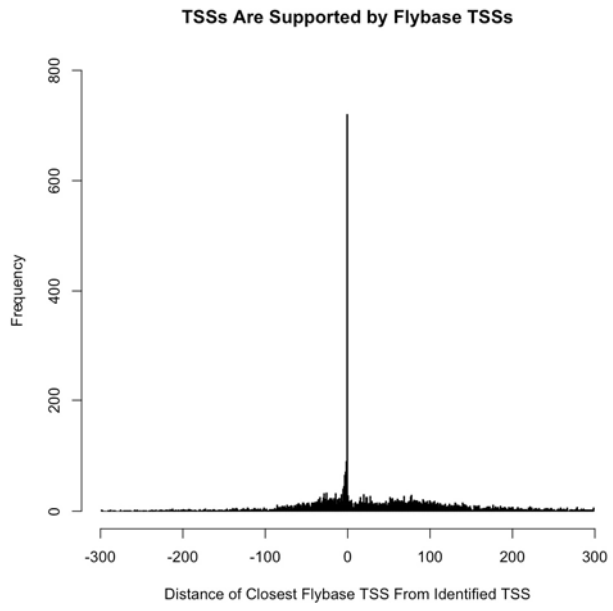


Figure 28: Flybase Location Differences

All TSSs in Flybase that were upstream of the most downstream start codon, and did not map to a start codon location, were selected for comparison. Each of the TSSs identified by the hierarchical clustering strategy was compared to all of the Flybase TSSs listed for the same gene. The smallest difference in location between the Flybase TSS and the selected TSS was calculated at 1bp resolution

using the selected TSS as a reference point, 0. The orientation of transcription of each gene was used to determine the orientation of the differences. A negative difference corresponded to a Flybase TSS being located upstream of the selected TSS, and a positive value signified that the Flybase TSS was downstream of the selected TSS. The plot covers a region of +/-300 nt, which covered 79% (4,406) of TSS matching to Flybase start sites. Compared to EPD, differences in start site locations are thus one order of magnitude larger at roughly the same coverage.

Akin to EPD, Flybase TSS locations showed a high percentage of overlap to the identified TSS locations (Figure 28), with 720 Flybase TSSs mapping perfectly to identified TSS locations; 30% located within 20bp; and 54% less than 100bp from identified TSSs. This resulted in the mean location of Flybase TSSs 98bp upstream of identified TSSs. Only 148 Flybase TSSs (2.6%) were more than 1,000bp from the identified TSSs. Of the 148 sites, only 13 Flybase TSSs were upstream of ours, and the remaining 135 Flybase TSSs were located more than 1,000bp downstream of identified TSSs. They may be genuine alternative downstream TSSs with too little support to be included in our set, or incomplete transcripts. The differences in TSS locations in Flybase are likely to stem from the application of various criteria to annotate transcripts and their 5′ ends, rather than the application of one consistent method for TSS calling.

The mean location of Flybase 5′ ends falling ~100 nt upstream of ours most likely reflects the long-standing strategy to call the most 5′ location of any single piece of evidence as the 5′ end of the whole transcript, even if many ESTs align to a more downstream location. While this strategy certainly leads to correct calls in terms of the genomic span of a gene, it does not correspond to high-quality TSS calls. We recognize

that experimental bias from PCR amplification may exist in the set of the most highly utilized TSSs (Ma et al. 2006), however, we believe it to be minimal, as the majority of differences between our TSSs and those in EPD and Flybase is less than 100bp. Such small differences in transcript size do not have a big impact on the rate of PCR amplification. As such, many analyses of CAGE data (Ponjavic et al. 2006) now commonly define the most frequent tag location as TSS.

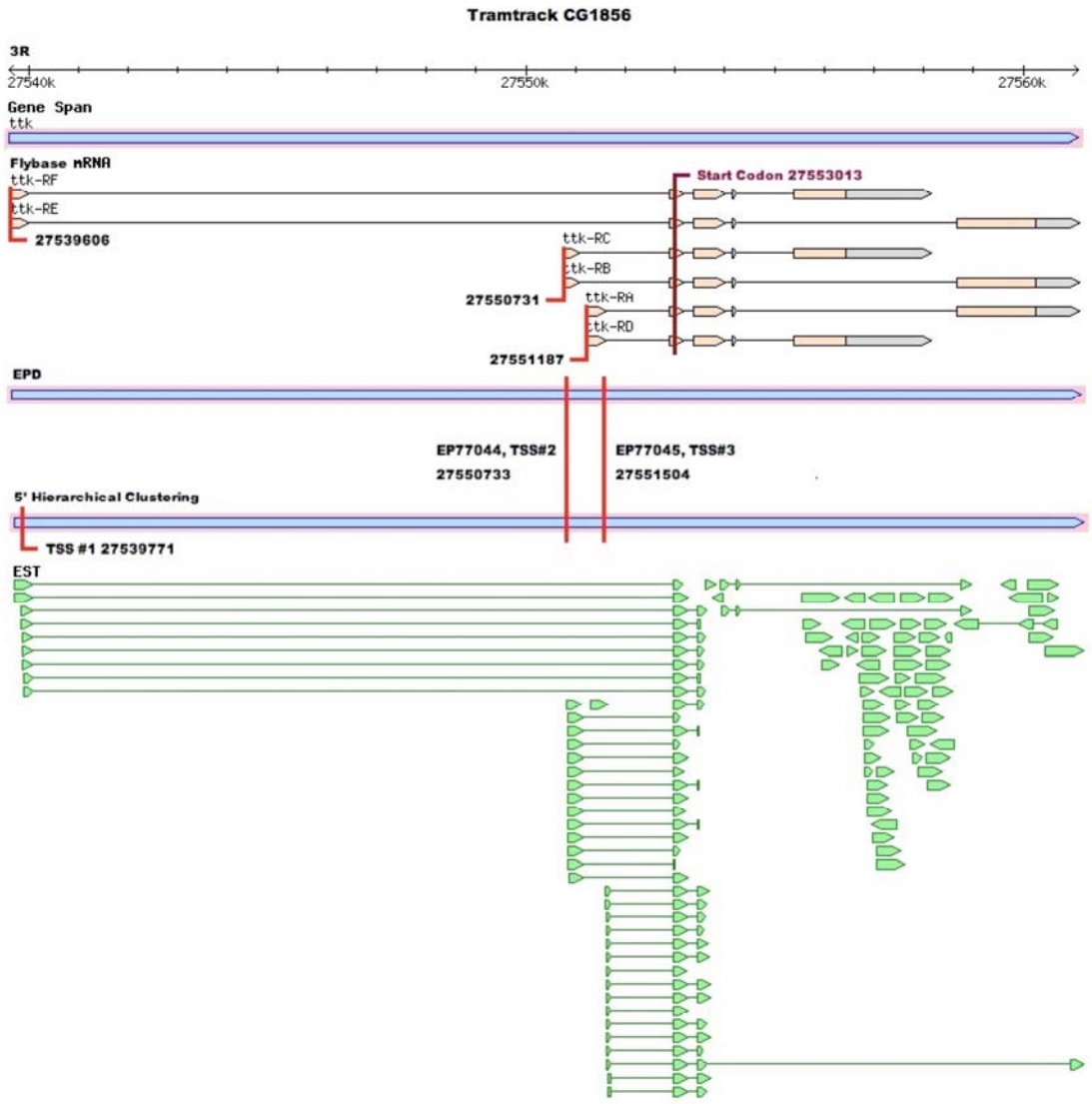


Figure 29: Alternative TSS Annotation for the Example Gene *Tramtrack*

Flybase annotation of TSSs at the tramtrack locus of Release 4.3 (Wilson, Goodman, and Strelets 2008). The gene span, Flybase mRNA, EST, and cDNA alignments were created using Gbrowse. The locations of the EPD sites, hierarchically clustered TSSs, and start codon were added manually.

While EPD and Flybase provide high quality support for the identified sites across the *Drosophila* genome, for a single gene, the TSS location information is often

incomplete using either database, and inconsistent using both. We illustrate this for the gene *tramtrack* (*ttk; CG1856*), a transcriptional repressor located on Chromosome 3R (Figure 29). There were three peaked TSSs listed in Flybase at locations 27539606 (TSS #1), 27550731 (TSS #2), and 27551187 (TSS #3). A fourth site at position 27552854 was listed, and is not shown, as it corresponded to the first nucleotide of the exon containing the start codon across all transcripts, and is likely to be an annotation artifact. The first TSS in EPD, EP77044, is 2 bp downstream of the Flybase TSS #2 at location 27550733. The second TSS, EP77045, occurred at location 27551504, and is 317bp downstream of Flybase TSS #3. The distributions of ESTs at both locations were classified as single initiation sites by EPD on account of their high frequency and small dispersion. In the hierarchically clustered set, we observed TSSs at locations 27539771 (TSS #1), 27550733 (TSS #2), and 27551504 (TSS #3). The two most downstream TSSs correspond to the TSSs in EPD, and the most upstream TSS is close to the first TSS annotated in Flybase, but missing in EPD. This agreement with EPD resulted from our use of a similar dataset and identification strategy. All three Flybase TSSs for *tramtrack* are upstream of TSSs in the EPD and our sets, highlighting the bias in the usage of the most 5′ evidence as TSSs, rather than the most highly utilized locations. The TSSs identified by hierarchical clustering thus supplement current annotations by providing precise and consistent TSS locations.

3.3.3.2 Core Promoters of Identified TSSs Contain Higher Frequencies of Essential Initiation Motifs Than the Most 5′ Sites in Flybase

To further demonstrate the reliability of the identified TSSs, we examined the presence of reported core promoter elements. For the Flybase motif comparison, we determined the frequency of the TATA, INR, DPE, and MTE motifs that are known to occur at a narrow distance from the TSS, in their preferred 10nt windows in the promoters. In addition, we analyzed the occurrence of the Ohler 1, DRE, Ohler 6, and Ohler 7 motifs in their 20bp preferred windows as defined in (FitzGerald et al. 2006) (Appendix B), as they are known to occur more widely throughout the core promoter. To allow for a clear comparison, we contrasted the number of motif matches in core promoters of genes with single promoters in our set to the number of matches in the core promoters of the most 5′ TSS coordinates defined for the same genes in Flybase.

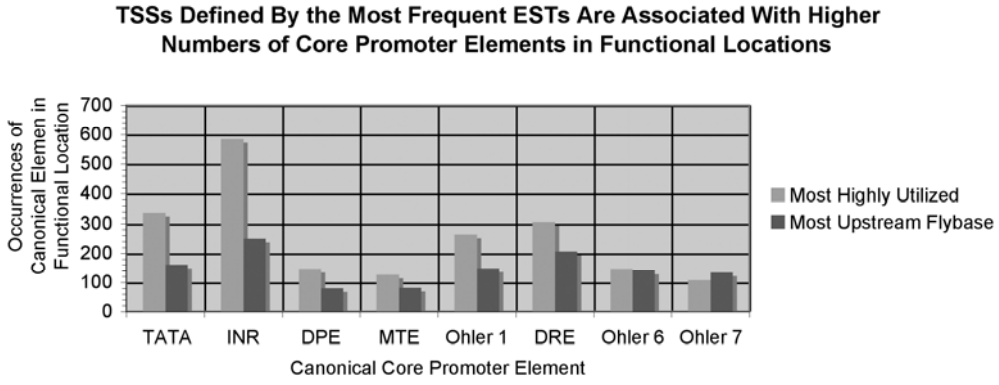


Figure 30: Presence of Core Promoter Elements

For 2,725 genes with exactly one TSS in our set and an annotated initiation site in Flybase, motif matches were identified in the preferred windows in their core promoter sequences using separate zero order Markov models as background.

There is a consistently higher number of motif matches in the promoters of the TSSs identified here, compared to those of the TSSs from the Flybase 5′ end annotations.

Figure 30 shows that each of the four location specific canonical elements was ~1.6-2.4 times more frequent in the core promoters of most highly utilized TSSs, than in those surrounding the most 5′ sites in Flybase. For the more broadly occurring Ohler 1 and DRE, a similar trend was observed with 1.8 and 1.5 times more occurrences of motif matches in the core promoters of the most highly utilized sites, respectively. For Ohler 6 and Ohler 7, the differences in motif frequencies were not distinctive; however, these two motifs have little location specificity (Figure 30), and the small differences in occurrences most likely reflect the breadth of locations of motif instances. The pattern that Flybase-derived promoters show a significantly lower number of known regulatory motifs has also been reported elsewhere (Berendzen et al. 2006). Due to the stringent window sizes, motif frequencies were overall lower than in some previous estimates; for instance, following the criteria in (Ohler et al. 2002), we detected an INR in 25% and a TATA box in 16.5% of promoters of genes with one TSS, which compares well with the presence of an INR in 26.3% and a TATA box in 19.3% of *melanogaster* promoters in the earlier study.

Looking at the presence of sequence motifs within *tramtrack* peaked promoters, an INR was present at both TSS#1 and TSS#3 as defined in our set, strengthening our

assignments for these TSSs, in spite of their considerably different locations in Flybase (Figure 29).

3.3.4 Sequence Elements are Associated With Different Initiation Patterns

For more than 20 years, it has been known that some promoters are highly position specific, while others are spread over larger regions (Bucher and Trifonov 1986). The analysis of large-scale CAGE data in mammals has confirmed the presence of peaked and broad promoters as a general phenomenon, and lead to a more precise definition of four different promoter shapes reflecting different initiation patterns (Carninci et al. 2006): (1) Single-peaked or focused; (2) broad or dispersed; (3) multimodal; and (4) broad with peaked(s). In the clustering analysis above, we identified two types of promoters: "peaked " for single TSSs, and "broad" for TSS cluster groups. The scale of the available fly data does not allow for a more precise sub-classification, but the two groups resemble the categories found in mammals to some extent, with the broad promoters being a potential combination of the categories (2) to (4).

Compared to mammals, analyses of the *Drosophila* genome have identified a larger set of sequence motifs enriched in core promoters. Ohler *et al.* predicted a set of ten motifs in the [-60,+40] bp region surrounding the TSS (Ohler et al. 2002); Fitzgerald et al. (FitzGerald et al. 2006) later identified 13 motifs with enrichment in the same region,

including nine of the ten motifs from Ohler *et al.* This knowledge allowed us to investigate whether the peaked and broad promoters were associated to specific core promoter elements, similar to the TATA box and CpG island biases found in mammals (Carninci et al. 2006). We focused on eight of the ten motifs in Ohler *et al.* that have either been biologically validated or previously reported as building blocks for core promoter sequence modules. The eight motifs included four location specific canonical motifs (TATA, INR, DPE, and MTE) (Juven-Gershon, Cheng, and Kadonaga 2006), and four motifs that have weaker positional biases, but were found to frequently co-occur in a specific order and orientation (Ohler 1, DRE, Ohler 6, and Ohler 7) (FitzGerald et al. 2006; Ohler 2006). Of the latter, only the role of the DRE in the recruitment of the polymerase has been unraveled (Hochheimer et al. 2002). We evaluated the occurrence of these eight motifs and their most frequently occurring modules in the core promoters surrounding 3,788 TSSs and 876 TSS cluster groups (see Materials and Methods). Because there were far more peaked promoters than broad promoters, their core promoters covered a three times larger genomic region. To provide an equal measure across both sets, and across motifs with differences in location preferences, motif matches were counted anywhere in the promoters, and the numbers of motifs found were then normalized to the number of occurrences per 100kb. For an estimation of the numbers of motif frequencies expected by chance, the analysis was repeated on three sets of 100bp regions surrounding randomly selected intergenic sites.

Figure 31A shows a clear separation in core element usage between peaked and broad promoters. While the TATA, INR, DPE, and MTE were more prevalent in peaked promoters, broad promoters had larger numbers of the Ohler 1, DRE, Ohler 6 and Ohler 7. As the TATA, INR, DPE, and MTE occur more frequently at specific locations from the site of initiation, and the Ohler 1, DRE, Ohler 6 and Ohler 7 have a weaker positional bias, peaked and broad initiation patterns directly correspond to the strength of location biases of the promoter elements that define them. With the exception of the INR, there were fewer occurrences of the location specific canonical elements in peaked promoters than there were of the motifs without location bias in the broad promoters. As this relationship appears after normalization, this suggests that the density of motifs is not linearly proportional to the genomic span of the core promoters, but rather that broad promoters, which include multiple closely spaced initiation sites, also contain higher densities of their most frequent elements.

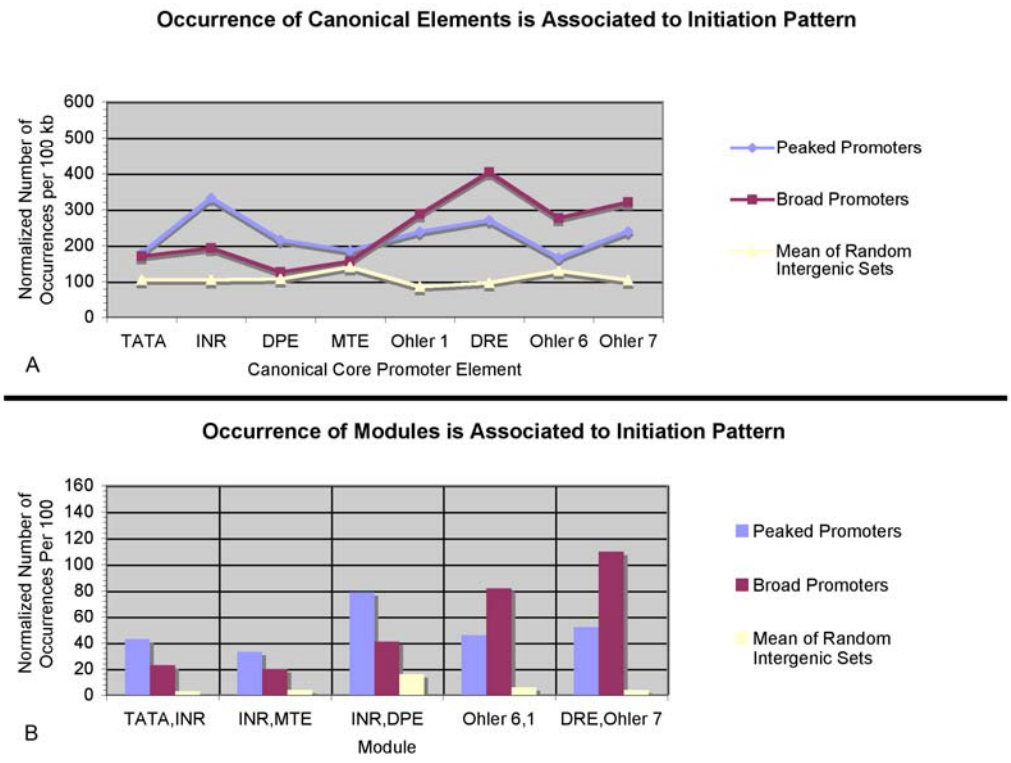


Figure 31: Core Promoter Elements Are Associated to Initiation Pattern

PATSER was used to evaluate the presence of the eight core promoter elements at any location in the 100bp sequences surrounding 3,788 TSSs, 876 TSS cluster groups, and three sets of 1,299 random intergenic sites. All counts were rounded to the nearest whole number after normalization. **(A)** Individual Motif Occurrences The number of motif matches were counted and normalized to the number of occurrences per 100kb. For the random intergenic sites, the mean numbers of motif occurrences across all three sets are shown. **(B)** Module Occurrences The number of pairs of motif matches present in the designated order, with respect to the orientation of transcription, were counted and normalized to the number of occurrences per 100kb.

The greatest difference in element frequency between peaked and broad promoters was observed for the INR and DRE. This suggests that the DRE may be of equal importance to the transcription for broad promoters as the INR is for the peaked

promoters. All motif observations were higher than the mean number of occurrences found across the three random intergenic sets, and random occurrence rates corresponded well to the expectation based on motif score cutoffs. When motifs in peaked promoters were constrained to their functional locations (see Materials and Methods), the same trends of occurrences were observed (Figure 32). We did not analyze restricted motif locations for the broad promoters, as multiple TSS reference points in the TSS cluster groups prevented distinct assignments within the overlapping core promoters.

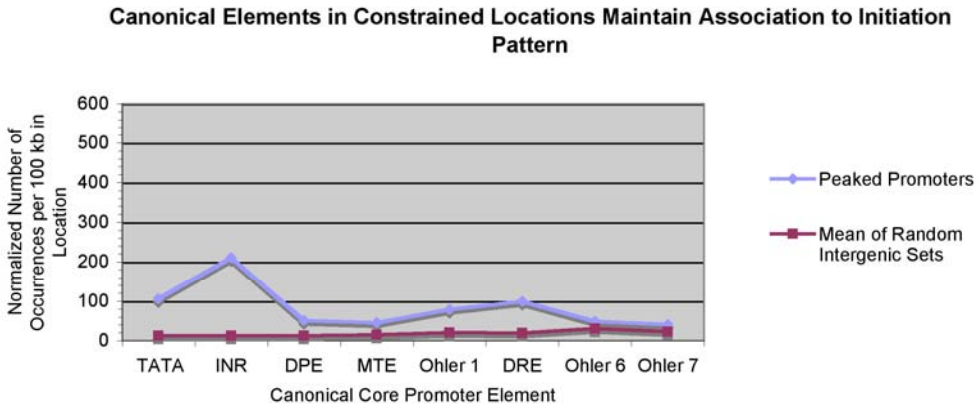


Figure 32: Sequence Elements in Preferred Windows of Peaked Promoters Preserve Trends of Motif Associations

Motif matches were constrained to their preferred windows in peaked core promoters and normalized to the number of occurrences per 100kb (see Materials and Methods). The mean number of occurrences across the three random intergenic sets is shown.

Next, we evaluated the presence of combinations, or modules, of known elements in the core promoters of the peaked TSSs and broad TSS cluster groups. A

previous study had identified five different core promoter modules, which we evaluated here: TATA/INR, INR/MTE, INR/DPE, Ohler 6/1, and Ohler 7/DRE (Ohler 2006). Overall, individual motifs had a much higher frequency in both peaked and broad *Drosophila* core promoters than observed collectively as part of motif modules (Figure 31). This may be evidence of motifs functioning independently of each other, in spite of their ability to synergistically cooperate. It may also suggest that individual motifs can have a general role in the binding of transcription factors to increase the overall rate of initiation, even though they may have a more restricted function when present in specific modules. Furthermore, as different repertoires of transcription factors are present under varying conditions, dual roles of motifs may correspond to different conditions. For instance, the TF binding to the DRE may be present individually in one condition, resulting in the DRE generally increasing the rate of transcription, while both TFs binding to the DRE and Ohler 7 may be present in a second condition, allowing for the complex to be more restrictive in interactions to recruit RNA pol II to the DNA.

Figure 31B shows that the TATA/INR, INR/MTE, and INR/DPE modules occurred more frequently in the peaked promoters, and the Ohler 6/1 and Ohler 7/DRE modules were more prevalent in the broad promoters. This corresponds with our results of the occurrences of the individual elements. It also shows that even though the Ohler 6 and Ohler 7 elements have a lower positional bias, they occur in a specific order within binding modules. All module occurrences in peaked and broad promoters were

far above the mean number found in the three random intergenic sets, although higher numbers of the most frequent modules appeared in the broad promoters than in those of peaked. This reaffirms that the broad core promoters of TSS cluster groups have a higher density of the most frequent modules of motifs than those of individual TSSs. Extending the analysis to three elements is limited by the rareness of such events, but analyses indicated that INR/MTE/DPE and TATA/INR/DPE occurred more often than triplets of elements with less positional bias (data not shown).

Finally, peaked core promoters were found to have higher frequencies of G (0.229) and C (0.234) than broad core promoters (G: 0.211 and C: 0.224) and the 100bp sequences surrounding the random intergenic sites (G: 0.203 and C 0.205). These results confirm previous work showing that core promoters with the DPE, INR, and TATA/INR have a moderate GC content, and core promoters with the DRE, and Ohler 1/6 elements have a GC-poor profile (Ohler 2006). With this analysis, we show that the GC content is not only characteristic of core promoter elements, but also of initiation patterns of transcription.

3.3.5 Conservation of Sequence Elements Differs Across Initiation Patterns

Given the different associations of motifs with initiation patterns, we sought to examine whether there were differences in the conservation of core promoter motifs across the 12 fully sequenced *Drosophila* genomes. We selected the promoters of

individual TSSs and TSSs in TSS cluster groups that had aligned sequences in all 12 species (see Materials and methods). This led to a reduced set of 4,243 promoters for 3,175 genes: 2,886 peaked TSSs, and 1,357 TSSs in broad promoters. We compared the conservation of the eight core promoter motifs in *D. melanogaster* to the other eleven genomes in a pairwise fashion (see Materials and methods). In other words, we assessed whether a presumably functional motif, defined by the occurrence of a motif match in the preferred window relative to the location of a mapped TSS in *D. melanogaster*, was still detected in a second species in the corresponding position in the alignment. Figure 33A shows that conservation levels of the INR motif ranged from approximately 90 to 95% for promoters in the *melanogaster* subgroup to approximately 50% for promoters in distantly related species. These levels directly correlate with the phylogenetic distances of the 12 genomes (Clark et al. 2007). Similar patterns are found for the other position-specific motifs, with the TATA box showing the highest level of conservation, and the MTE the lowest in more distant species. For the other four motifs, the conservation levels were consistently lower.

While this analysis showed clear trends, it did not indicate whether such observations could arise from chance. We therefore determined the fraction of pairwise conserved motif matches by dividing the number of conserved motif instances in the preferred window over the total number of occurrences anywhere in the *D. melanogaster* promoters. After repeating this analysis on a set of similar sized random intergenic

sequences, we took the ratio between promoters and random sequences as the motif enrichment score; for *D. melanogaster* alone, this score simply indicated the enrichment of hits in the preferred window (Figure 33B). In general, ratios were higher for the position-specific motifs INR, TATA, MTE, and DPE, with the INR exceeding enrichments of 30-fold. While there was a lower but consistent score for Ohler 1 and DRE, the motifs Ohler 6 and Ohler 7 did not clearly exceed a ratio of 1 in *D. melanogaster*, indicating that the preferred windows taken from (FitzGerald et al. 2006) were not actually enriched above background. The total number of conserved instances was quite low for these motifs, and the higher scores seen for more distantly related species may be regarded with caution, as they could simply be a side effect of the small sample size. Nonetheless, we saw that the motifs that were less restricted in their relative location to the TSS showed a lower level of conservation in the aligned locations.

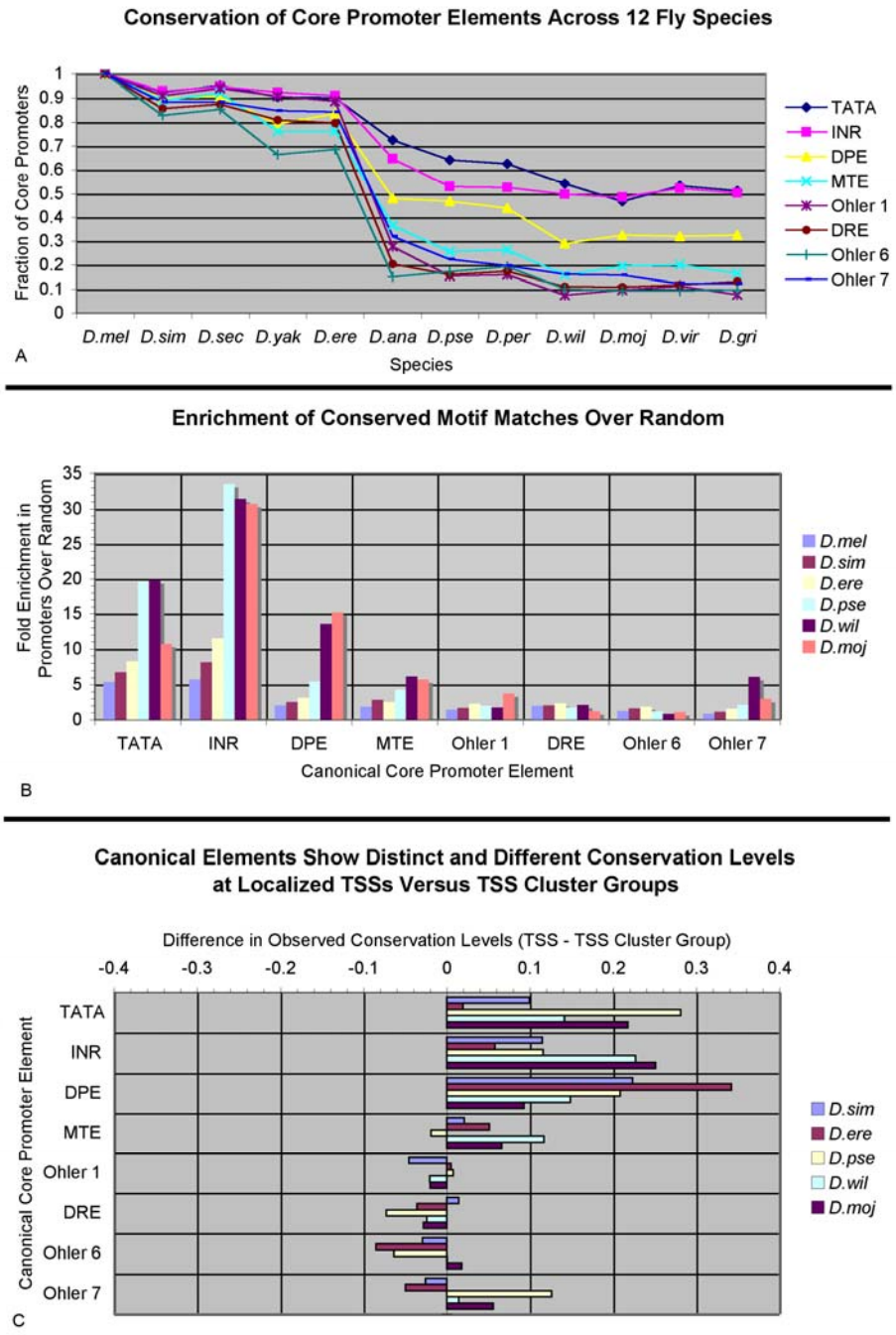


Figure 33: Evolutionary conservation of sequence elements.

The core promoter sequences surrounding each *D. melanogaster* TSS were mapped to orthologous locations in the 12 *Drosophila* genomes. **(A)** Conservation of sequence elements across the 12 fruit fly genoomes. The set of *D. melanogaster*

promoters having an element present in its preferred window was selected, and the fraction of all orthologous sequences with the motif present was assessed in a pairwise fashion with the other 11 species. The figure indicates a sharp decline in the conservation of the elements outside of the melanogaster subgroup. **(B)** Enrichment of conserved motif matches in promoters over random sequences. The plot shows the fold enrichment of the fraction of total *D. melanogaster* motif matches conserved in the preferred window of 100-bp sequences surrounding detected TSSs compared to random intergenic locations. For clarity, the plot shows only five out of the eleven species in the total pairwise comparisons. **(C)** Differences in conservation of canonical elements between peaked versus broad promoters. After splitting the motif matches used in (a) by their occurrence in peaked versus broad promoters, there are noticeable differences between the conservation levels of motifs. For clarity, we again only show five out of the eleven pairwise species comparisons. *D.mel*, *D. melanogaster*; *D.sim*, *D. simulans*; *D.sec*, *D. sechellia*; *D.yak*, *D. yakuba*; *D.ere*, *D. erecta*; *D.ana*, *D. ananassae*; *D.pse*, *D. pseudoobscura*; *D.per*, *D. persimilis*; *D.wil*, *D. willistoni*; *D.moj*, *D. mojavensis*; *D.vir*, *D. virilis*; *D.gri*, *D. grimshawi*.

Given that these two motif sets were shown to be associated with different initiation patterns, we assessed whether motifs in peaked promoters exhibited different conservation patterns than those in broad promoters. Figure 33C shows that there are indeed strong differences in the conservation levels of motifs across initiation patterns. Conservation levels of localized motifs (TATA, INR, DPE, MTE) were consistently higher when they occurred at peaked TSSs versus TSSs in broad promoters. This trend was mirrored in a somewhat weaker fashion by the set of motifs with lower positional preference (Ohler 1, DRE, Ohler 6, Ohler 7), which were more conserved in peaked than broad promoters. Observations on promoter conservation and TSS turnover have been reported for human-mouse comparisons supported by 5′ capped tag data (Frith et al. 2006). In particular, findings indicated that some alternative promoters experience a

lower negative selective pressure, and this may reflect an intermediary stage of a TSS turnover event. Our findings here indicate that selective pressure on the motifs in promoters also depends on the initiation patterns, with evidence that broad promoters may experience more frequent functional motif turnover due to the lowered restrictions on relative spacing of enriched motifs, and/or the presence of other functional promoters in the close vicinity.

Looking at the conservation of motifs for the *ttk* case study (Figure 29), we recall that two INR motifs were present in the preferred location of the peaked promoters of TSS#1 and TSS#3. The initiator motif in the TSS#1 promoter was conserved across all 12 species, and the initiator in the TSS#3 promoter was conserved within the 5 species of the *melanogaster* subgroup. This illustrates the existence of differences in motif occurrence and conservation levels at alternative start sites.

3.4 Discussion

The identification of 5,665 TSSs from hierarchical 5′ EST clustering provide a comprehensive map of reliable transcription start sites in *D. melanogaster*. By designating TSS positions at the location of the highest EST frequencies within a clearly delineated cluster, instead of at every mapped location (Carninci et al. 2005) or the most 5′ one (Zhang and Dietrich Identification and characterization of upstream open reading frames (uorf) in the 5' untranslated regions (utr) of genes in saccharomyces cerevisiae

2005), we were able to gain new insights on the architecture of core promoters and their associations to conditions. The two most sensitive parameters in the clustering algorithm were the standard deviation and minimum frequency. We selected informed values based on previous analyses of *Drosophila* core promoters, however, increasing or decreasing these values changes the number of TSSs identified. The saturation of the *D. melanogaster* transcriptome by the current set of sequence tags is certainly incomplete, and additional TSSs exist beyond the high quality set identified in this work. Nevertheless, the TSS map provided here should serve as a useful resource for information regarding condition-specific transcription initiation, and for computational modeling of promoter regions.

Given the amount of available data and correspondingly chosen clustering parameters, all TSS positions are separated by at least 20bp, with the consequence that motif assignments, which were restricted to small, preferred windows relative to the TSSs, could be made to individual sites. In mammalian studies on large-scale 5'capped transcript datasets, initiation sites were observed at many closely spaced locations and called at single-nucleotide resolution (Ponjavic et al. 2006). While some of the initiation frequencies at this resolution have been shown to be condition-specific, broader TSS initiation patterns may potentially be a result of some degree of sloppiness in the transcriptional machinery, and functional consequences of such differences on transcription are of yet unclear.

Drosophila core promoters distinguish themselves from other eukaryotic species investigated so far, by being defined by a repertoire of well-known sequence motifs. A concurrent recent study explored how promoters relate to one another across alternative TSSs and adjacent genes (Zhu and Halfon 2009). Here, we examined differences in motif frequencies of peaked and broad promoters. We showed that peaked promoters have higher frequencies of the location specific motifs (TATA, INR, DPE, MTE) and their corresponding modules TATA/INR, INR/DPE, INR/MTE, and a higher GC content. The importance of the location of elements in peaked promoters with respect to the TSS may reflect the binding architecture of specific TAFs in the RNA pol II.

While the core promoters of broad promoters showed an increase in the less location enriched elements (Ohler 1, DRE, Ohler 6, Ohler 7), their modules Ohler 6/1, DRE/ Ohler 7, and had a poor GC profile (Ohler 2006). While this segregation of sequence elements in the core promoters of TSSs and TSS cluster groups is striking, it is not complete. Nearby peaked promoters may be designated as broad promoters, and vice versa, as a promoter may appear peaked due to a limited amount of data.

Our findings suggest that the core promoters of peaked TSSs in *Drosophila* are functionally equivalent to those surrounding the single dominant peaked TSSs in vertebrates. The peaked promoters in both *D.melanogaster* and vertebrates have single, well-defined sites of initiation, contain location specific motifs, and are associated to similar functional subsets of genes. Similarly, we showed that the core promoters of

broad TSS cluster groups in *Drosophila* are functionally equivalent to the broad regions of initiation in vertebrates (Carninci et al. 2006). Both are composed of multiple initiation sites, with no fixed spacing between them, contain motifs without a location enrichment, are void of the location specific motifs, such as the TATA, and are also present in similar functional subsets of genes. It is important to recognize, however, that we are comparing the functional usage of each 'type' of core promoter across *Drosophila* and vertebrates, and not the actual sequence features that comprise them; *Drosophila* and vertebrates have core promoter sequence features that are uniquely adapted to the transcription initiation machinery of each species. For instance, out of the eight motifs used in this study, only three motifs (TATA, INR, and DPE) have been shown to be functionally relevant for transcription initiation in vertebrates (FitzGerald et al. 2006). In turn, other sequence elements such as the downstream element DCE play an important role in human (Lee et al. 2005). In our analysis, broad promoters were found to contain higher densities of the most frequent motifs and modules. TSS cluster groups may be 'hotspots' in animal genomes and have higher probabilities of gaining additional motifs and modules. As they define larger domains, broad promoters may be susceptible to higher probabilities of gaining motifs and modules. It will be interesting to explore whether, similar to other genomic properties including gene family sizes (Rach 2004) and protein folds (Koonin, Wolf, and Karev 2002), the relationship between motif density and genomic span of initiation is be scale free.

The most salient difference between fruit fly and vertebrate promoters regards the presence of CpG islands. In vertebrates, CpG islands are characteristic of broad initiation regions, and are less frequent in peaked promoters, while in *D. melanogaster*, CpG islands do not exist, and peaked promoters have higher frequencies of G and C than those of broad promoters. This may indicate that the shape of promoters may be independent of the functional properties of CpG islands. The core promoter motifs may have been decoupled from CpG islands, or the properties of CpG methylation, selectively in the evolutionary history of *D. melanogaster*, as many other insect taxa have CpG methylation and orthologous proteins that catalyze it in vertebrates (Tweedie et al. 1999; Wang et al. 2006). Furthermore, the core promoter motifs may be more dependent on the epigenetic features of the genome, such as the organization of histones and histone methylation, rather than on the properties of the DNA sequence itself.

Overall, individual motifs had a much higher frequency in both types of *Drosophila* core promoters than motif modules. This may be evidence of motifs functioning independently of each other, in spite of their ability to synergistically cooperate. It may also suggest that individual motifs can have a general role in the binding of transcription factors to increase the overall rate of initiation, even though they may have a more restricted function when present in specific modules. Furthermore, as different repertoires of transcription factors are present under varying conditions, dual roles of motifs may correspond to different conditions. For instance, the

TF binding to the DRE may be present individually in one condition, resulting in the DRE generally increasing the rate of transcription, while both TFs binding to the DRE and Ohler 7 may be present in a second condition, allowing for the complex to be more restrictive in interactions to recruit RNA pol II to the DNA.

While we did not explore this in the current study, a different prominent effect of alternative promoters lies in downstream effects: the diversification of the gene's isoforms, an increase in the complexity of the gene's architecture and possibly, an expansion of the biochemical role of the gene's function. Analysis on full-length mouse cDNAs has shown a significantly higher correlation between the utilization of alternative promoters and the occurrence of alternative splice isoforms and multiple start codons, than for genes with one promoter (Zavolan et al. 2003). The association of alternative promoters to alternative transcript isoforms may simply result from the condition-specific expression of transcription and splice factors that independently lead to diversified isoforms. A more intriguing model suggests a direct link between the preferential recruitment of splicing factors to mRNAs transcribed from alternative promoters and the recruitment of splicing factors by condition and promoter-specific transcription factors (Chern et al. 2008; Cramer et al. 1999; Gendra et al. 2007). This could result from multiple TSSs always generating different 5'UTRs sequences, which are known to harbor functional elements. A third model suggests that alternative promoters

have more subtle effects, such as affecting the mRNA's half-life, rate of translation, localization, or overall level of protein production.

Our study provided a high-quality data set to assess the conservation of core promoter elements across the recently published 12 *Drosophila* genomes. As we have experimental data for one species, we can only evaluate the loss of a *D. melanogaster* site in the corresponding location in another species. The fraction of candidates with non-conserved promoter elements in the *melanogaster* subgroup (approximately 10% depending on the motif and species) agrees with the turnover frequency measured by the ChIP-validated Zeste binding site (Moses et al. 2006). The observed conservation levels drop drastically outside the *melanogaster* subgroup. A larger evolutionary effect in more distal species is certainly expected, but the recently observed low performance of multiple alignment algorithms on distal non-coding regions is likely to be a strong contributor to this observation (Huang, Nevins, and Ohler 2007; Pollard et al. 2006). Promoters of alternative TSSs, in particular those of broad TSS cluster groups, show a distinctly lower level of conservation of motifs across the 12 *Drosophila* genomes. This provides initial evidence of an average lower negative selective pressure on alternative and broad promoters, linked to the presence of functional motifs. A possible explanation for this effect was given in a recent TSS study on human and mouse, by using high-throughput CAGE sequence tags (Tsuritani et al. 2007). This study showed that alternative TSSs may arise in an intermediate stage of the process of TSS turnover. In

support of this, an analysis of primate core promoters gave evidence for accelerated substitution rates (Liang, Lin, and Li 2008).

TSSs may be more dynamic than previously thought (Gross and Oelgeschlager 2006). In addition to the effects discussed above, they are involved in enhancer functionality (Butler and Kadonaga 2001; Ohtsuki and Levine 1998), transcriptional interference (Martens, Laprade, and Winston 2004), condition restricted TAF utilization (Hiller et al. 2004), and the maintenance of Internal Ribosome Entry Sites (IRESes) (Hernandez et al. 2004; Vazquez-Pianzola et al. 2007). As the amount of data increases from capturing 4,000 genes in this study to the 13,767 genes present in the *D. melanogaster* genome, we expect the number of genes with alternative TSSs to scale accordingly. The first sets of 5'capped high-throughput transcript data have become available concurrently to our study, and such data will provide the necessary scale to follow up on our observations (Ahsan et al. 2009). Long underestimated in importance, the utilization of TSSs has now been shown to contribute significantly to the complex regulatory code of the eukaryotic transcriptome.

4. Conditions Specificity of Single and Alternative Promoters

Elizabeth Rach conceived and performed all of the work in this chapter, except for the GO analysis of genes with alternative TSSs having distinct condition associations, which was contributed by William Majoros from Dr. Uwe Ohler's lab. The work was published in *Genome Biology* in July 2009.

4.1 Introduction

A wide range of animal genes possess clearly separated alternative promoters that are associated with specific functional consequences (Davuluri et al. 2008). In *D.melanogaster*, several well-known genes are known to use well-separated alternative promoters under different conditions. For instance, the transcriptional activator Hunchback (*Hb*) has two isoforms with different maternal (distal promoter) and zygotic (proximal promoter) patterns of initiation (Margolis et al. 1995; Margolis et al. 1994). Alcohol Dehydrogenase (*Adh*) utilizes two promoters, one during embryonic development and the second in adulthood (Corbin and Maniatis 1989). However, the extent to which such condition-specific variability is reflected in mammalian and *Drosophila* core promoters is so far mostly unclear.

As the presence and levels of TFs varies across tissues and time periods, arrangements of binding sites to which the TFs associate in the promoter region should

reflect, to a certain degree, the conditions under which a specific core promoter is utilized (Beckett 2001; Remenyi, Scholer, and Wilmanns 2004). Gene ontology (GO) and microarray analyses have proved valuable in associating individual core promoter elements to various functional terms, such as germline expression and in capturing general trends of sequence element enrichments for various tissues and the embryo and adult stages of the fruit fly life cycle (FitzGerald et al. 2006). In addition, recent studies on tissue specific TAFs showed that the core machinery is remodeled in specific conditions (Deato and Tjian 2007; Metcalf and Wassarman 2006), and it is expected that the specificity of TAFs is encoded in additional core promoter sequence elements. However, due to genome wide expression studies typically being based on gene-wide probes located in the coding or 3′ untranslated regions, the sequence elements governing this spatiotemporal regulation have been elusive. Expression patterns made on a whole gene basis, such as those in FlyAtlas (Chintapalli, Wang, and Dow 2007), and in various conditions (Spellman and Rubin 2002), have neglected differences in distinct transcript variants, and ultimately their core promoters. Low-throughput studies using primer extension or 5′ RACE to evaluate the utilization of promoters on a higher resolution have been typically done under one condition, which has further restricted possible conclusions about the condition specific usage of alternative promoters.

The recent high throughput sequencing efforts based on 5′ capping protocols have now generated CAGE tags for human and mouse under numerous conditions

(Carninci et al. 2005; Kimura et al. 2006; Valen et al. 2009). These efforts showed that initiation patterns of the 5′ end locations contained tags generated from mixtures of different conditions. This outcome greatly increased the complexity of characterizing the utilization of initiation sites. The array of conditions captured in the CAGE sets for mouse and human is not yet available across the 12 *Drosophila* genomes (Clark et al. 2007; Stark et al. 2007), but a moderate assortment of different conditions used to generate the 5′ capped ESTs exists (Celniker et al. 2002). In spite of this data, genome-wide efforts to assign TSSs to specific conditions, and to analyze associations of sequence elements and modules to spatiotemporal conditions, have been minimal in *Drosophila* and vertebrates.

Here, we identify distinct associations of TSSs to spatiotemporal conditions based on the Shannon entropy of EST frequencies from different libraries. We investigate the specificity of alternative promoter utilization at higher temporal resolution by using available expression data from tiling arrays during embryonic development. Lastly, we identify intriguing trends of core promoter elements and their corresponding modules in maternally and zygotically utilized sites. Our analysis demonstrates that sequence elements in core promoters are directly associated with the spatiotemporal conditions under which they are utilized.

4.2 Materials and Methods

4.2.1 Shannon Entropy to Measure Condition Enrichment

We assessed the condition association of TSSs by computing the Shannon entropy of the ESTs of each (sub-) cluster from which they were identified, using a protocol following previous methods (Schug et al. 2005). First, we defined $w(tss,i) = N(tss,i) / (x_i + 5{,}665)$ for (sub-)cluster *tss*, condition *i*, where $N(tss,i)$ = the number of ESTs in each (sub-)cluster *tss* and condition *i*, x_i = the number of ESTs for one condition across all (sub-)clusters, and $5{,}665$ = the total number of (sub-)clusters in the analysis. In other words, $w(tss,i)$ represents the normalized expression counts of the ESTs by condition and the overall size of the dataset. Next, we obtained the probability of observing an EST for each condition in a (sub-) cluster, $P(i \mid tss) = w(tss,i) / N_{tss}$, for N_{tss} = the total number of ESTs in the (sub-) cluster across all conditions. To avoid arbitrarily low entropy values, we smoothed the data for conditions with no ESTs by setting $P(i \mid tss) = .001$. We calculated the entropy $H_{tss} = - \sum P(i \mid tss) \log_2 P(i \mid tss)$ by summing across all conditions *i* for each *tss*. Then, we penalized entropy values to account for the disparity in sampling depth across conditions, $Q_{i,tss} = H_{tss} - \log_2 P(i \mid tss)$.

We characterized the condition utilization of each (sub-) cluster by using an EST frequency threshold and the penalized entropy values, $Q_{i,tss}$. Only (sub-) clusters having at least 3 ESTs from a condition were evaluated further to prevent potential false assignments due to a low frequency of ESTs. The entropy values for H_{tss} ranged from *0*

to $log_2(c)$, for c = the number of conditions. In our analysis, $c = 9$ (eight distinct conditions and one diverse condition), and values for $Q_{i,tss}$ ranged from 0 to $log2(9) - log2(.0001)$, or 16.458.

Q values naturally segregated into three clearly distinct groups (Figure 34). Entropy values close to zero signified (sub-) clusters with ESTs mainly from one condition. Larger entropy values characterized (sub-) clusters with ESTs that were more broadly distributed across libraries, but still mainly concentrated in one or two conditions. The greatest entropies denoted (sub-) clusters with ESTs spread across many of the eight conditions. On account of these groups, we classified the TSS associations into three categories (condition specific, condition supported, and mixed) based on chosen cutoffs of $Q_{i,tss}$. TSSs were declared condition specific if $0 \leq Q_{i,tss} \leq 1$, and there were less than two ESTs from other conditions, and condition supported if $0 \leq Q_{i,tss} \leq 1$, and more than two ESTs were generated from other conditions. We also classified TSSs as condition supported if $1 \leq Q_{i,tss} < 10$. TSSs with $Q_{i,tss} \geq 10$, and those that were classified as specific or supported by more than 2 of the 8 distinct conditions, were deemed to have mixed association. Finally, TSSs that were specific or supported by the diverse condition were assigned mixed association by default.

Furthermore, we validated the significance of the TSS associations by performing 100 random permutations on condition labels to (sub-) clusters. For each permutation, we preserved the same partitioning of EST frequencies across the (sub-) clusters as in the

identified set, and applied Shannon entropy to the (sub-) clusters to identify specific, supported, and mixed condition associations. We then fitted the total number of condition specific associations across the 100 permutations to a Gaussian distribution and used the Gaussian to obtain an empirical P value. To evaluate the statistical significance of the gene associations, we performed 100 random permutations of the pattern associations called for the identified set of 5,665 TSSs and fit a Gaussian to the number of genes with the same condtion. On average, 242 genes had the same condition associations across alternative TSSs, and 983 genes had different associations across alternative TSSs. The counts we observed in the real data significantly differed from these numbers ($p << .001$). This was also the case when we repeated the analysis on the random permutations of EST condition labels created when we evaluated the significance of the condition associations of individual TSSs. The same partitioning of TSSs per gene was preserved.

4.2.2 Evaluating Temporal Usage of Promoters by Affymetrix Tiling Arrays

Our analysis is based on a published embryonic time course (Manak et al. 2006). Fluorescence intensities were observed for 3,075,693 probes across 105,897,358 bp of the fly genome. Each of the oligos used in the array was 25bp in length, spaced at ~35bp intervals genome-wide. Unlike ESTs, which allowed us to assign TSS associations at the level of individual nucleotides, the limited tiling resolution restricted our ability to

distinguish differences in transcriptional activity of promoters at individual closely spaced TSSs. Therefore, we analyzed the temporal embryonic utilization of peaked promoters separated by more than 100bp and broad promoters. The spatiotemporal utilization of the most upstream TSS in a broad TSS cluster group was chosen to characterize the whole group. This resulted in 4,664 well-separated promoters.

Given this level of resolution, we needed to distinguish sites of true transcription initiation from background fluorescence noise. Transcription levels of internal coding regions of genes can be distinguished from background by large differences in fluorescence levels. However, boundaries between TSS locations and non-transcribed adjacent intergenic sequences are typically not as clear. Recent studies used a median estimator and the binomial distribution to distinguish expression from background (Bertone et al. 2004; Kampa et al. 2004; Royce et al. 2005). These methods effectively identified positive expression of individual tiles; however, they did not take into account neighboring tiles to determine expression boundaries. As we were interested in finding the boundaries of transcripts, we subtracted the median of fluorescence intensity of 3 tiles upstream of TSS locations from the median of fluorescence intensity of 3 tiles downstream, with respect to the orientation of transcription. Tiles containing the TSS location were excluded from the analysis because we did not expect such probes to show consistent expression. TSSs may not be detected in this analysis if intensity differences fell below the specific thresholds, or if genes had exceptionally short first

exons for which the tile size of 25bp is too large to obtain a reliable signal above the background.

Due to the differing levels of total transcription across the 12 two-hour periods, cutoffs were determined independently for each time point (Table 1). A mixture model of two Gaussians was fit to the differences of each time point using Expectation Maximization (EM). The point of intersection of the two Gaussians was rounded up to the nearest .5 and declared the threshold. All promoters having differences greater than the threshold were deemed transcribed (T) for that time point. Promoters having differences in median fluorescence intensity less than the time point specific threshold were declared non-transcribed (N).

Table 1: Affymetrix Cutoffs for Determining Significance

Time Point	Hours	Difference Threshold	Number of False Positives	False Positive Rate
1	0-2	36.5	103	0.022
2	2-4	23.5	94	0.020
3	4-6	21.0	158	0.034
4	6-8	23.5	121	0.026
5	8-10	29.5	122	0.026
6	10-12	35.0	126	0.027
7	12-14	24.5	113	0.024
8	14-16	32.0	162	0.035
9	16-18	28.0	113	0.024
10	18-20	21.0	149	0.032
11	20-22	23.0	92	0.020
12	22-24	18.5	93	0.020

The fraction of promoters transcribed at each time point was determined by dividing the number of transcribed promoters at each 2-hour period by the total number of promoters. A paired t-test was applied to the fractions of transcribed peaked versus broad promoters to evaluate statistical significance. The same strategy was used to compare the fraction of peaked versus broad promoters with embryo EST associations over all 12 time points, and to compare the total number of initiation sites with embryo EST associations to those without. For the evaluation of the association of both types of promoters to embryo and non-embryo ESTs associations, without the tiling array data, a χ^2 test with Yates' continuity correction was applied. A Bonferroni correction was used in all tests, reducing the effective significance level to .01.

To determine the expected fraction of false predictions at these cutoffs, we randomly selected 4,664 random intergenic sites as a control dataset. For each of these sites, we evaluated the difference in fluorescence intensities akin to that of the set of TSSs. We used the same threshold values, and assumed the sites had positive orientation.

4.2.3 Temporal Utilization of Core Promoter Motifs

In the core promoter analysis, maternally inherited sites were defined as having utilization during time points 1 and/or 2 in the tiling array. Sites with zygotic transcription were required to have utilization during at least one two-hour period from

4 through 12, and sites with both maternal and zygotic utilization needed to satisfy both requirements. The motif matches for the eight elements and their modules previously identified (see Materials and Methods Chapter 3) were summed up separately for these three sets. As the initiation pattern does not play a role for random intergenic sites, the mean numbers of elements identified in the 1,299 random sites served as a baseline.

4.3 Results

4.3.1 TSSs Have Distinct Associations With Conditions Derived From EST Libraries

4.3.1.1 TSS Patterns of Utilization

Sites of transcription initiation are determined by the conditions under which transcription factors mediate the recruitment of RNA polymerase II to the core promoter. Associations of TSSs to conditions can give insight into the utilization and organization of transcription factor binding sites. For this reason, we characterized the condition associations of the set of 5,665 TSSs identified from the hierarchical clustering of 5′ ESTs in *D. melanogaster.* As mentioned above, the cDNA library information for each of the ESTs was mapped to one of eight distinct conditions (embryo, larva/pupa, head, ovary, testes, Schneider cells, mbn2 hemocytic cells, and fat body) plus a default (diverse) category. Overall, the data was more descriptive of spatial body parts than of well-resolved temporal stages of *Drosophila* development. Then, we used Shannon entropy to evaluate the specificity of association of a TSS to one condition relative to the

other conditions (see Materials and Methods). Entropy had been used previously to determine the associations of genes to tissues in the mouse and human genomes from ESTs (Schug et al. 2005), and was a natural fit to the data. For each condition, the penalized entropy values (Q_{tss}) segregated into three groups that we used to classified as: condition specific, condition supported, or mixed (see Materials and Methods, Figure 34). This classification scheme provided an initial framework for characterizing TSS condition associations, and ultimately promoter utilization. It can be easily expanded to future data sets and applied to other types of count data. As current TSS condition associations were made given the available set of ESTs, they may change with the inclusion of additional 5′ capped data.

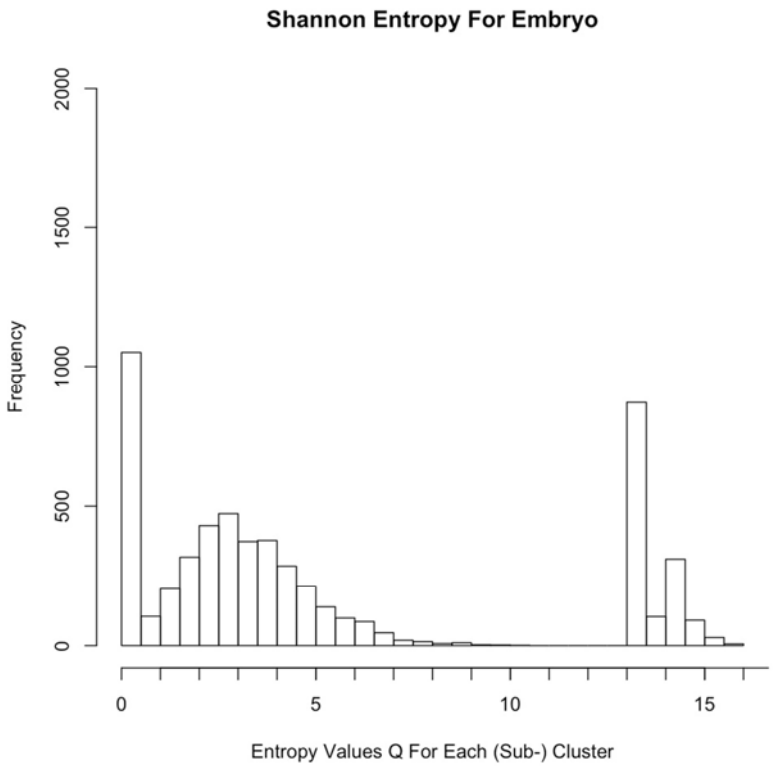


Figure 34: Shannon Entropy Values Segregate Into Three Groups

The distributions of ESTs in the (sub-) clusters used to call TSSs were evaluated using Shannon entropy. As example, the figure shows the entropy histogram for the embryonic condition with bins of size 0.5. The $Q_{Embryo,tss}$ values naturally separate into 3 groups: those less than 1, those between 1 and 10, and those greater than 10. The large frequency of $Q_{Embryo,tss}$ values between 13 and 13.5 results as an artifact of using .0001 to smooth p (i | tss) for (sub-) clusters containing ESTs mainly from one non-embryo library.

There were 1,997 (35%) TSSs with specific associations (Figure 35), and 1,612 (29%) TSSs with supported associations in one of the eight conditions (Appendix C). Together, almost two thirds of the TSSs had associations to only one condition. Specific and supported assignments existed for TSSs across all conditions, with the embryo and

the head having the largest numbers of specific or supported sites. The testes had the third largest number of specific TSSs (247), and the ovary had the smallest number of specific TSSs (9). The numbers of testes and ovary TSSs were comparatively higher than their fraction within the set of filtered ESTs. There were 14% of TSSs that were supported in two conditions. The two largest pairs of condition associations were embryo:head and embryo:Schneider cells. The embryo:head pair can be accounted for by the large sizes of the ESTs in their libraries, and the embryo:Schneider cell pair can be explained by the fact that Schneider cells are derived from embryos at 20-24 hours of development. There were 1,275 (22%) TSSs classified as having mixed associations. By default, we labeled TSSs that were specific or supported for the diverse condition as having mixed associations because their supporting ESTs were derived from broad or unknown conditions.

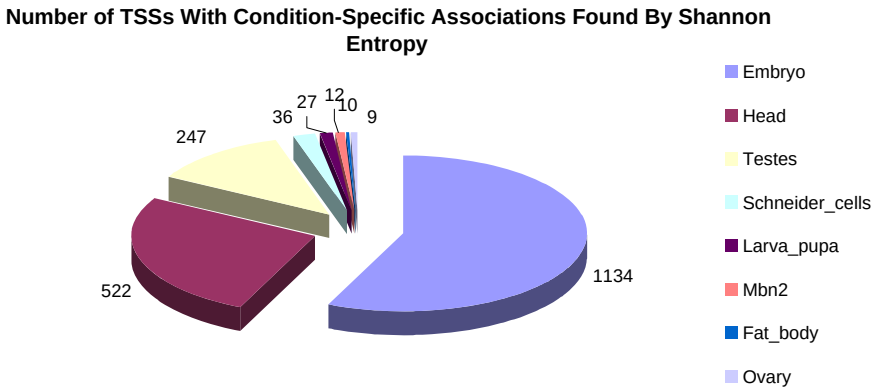


Figure 35: Condition Specific Associations For the Set of Identified TSSs As Determined by Shannon Entropy

Shannon entropy was applied to 72,535 ESTs in the (sub-) clusters of 5,665 identified TSSs. There were 33,077 ESTs from embryo, 23,361 from head, 3,903 from Schneider cells, 2,883 from testes, 2,267 from larva pupa, 1,978 from ovary, 699 from mbn2 cells, 471 from fat body, and 3,896 with the diverse label. The degree of association of the TSSs to the spatiotemporal conditions was evaluated using EST frequency, Shannon entropy, and a tripartite classification system (see Materials and Methods). The numbers of TSSs with specific associations are shown.

For the previously mentioned example gene *tramtrack,* all three TSSs had embryo associations. The two most upstream TSSs were embryo supported, and the third downstream TSS was embryo specific. The associations corresponded to the known expression of the gene during embryogenesis for various functions, including the regulation of proper development of tissues (Araujo, Cela, and Llimargas 2007) and the determination of cell-fate (Bardin, Le Borgne, and Schweisguth 2004). While these assignments do not determine function, they help to define the scope of alternative promoter utilization and contribute novel information about expression patterns.

4.3.1.2 Genes With Alternative TSSs

When considering condition associations on a gene level, the numbers of specific, supported, and mixed TSSs did not significantly differ for genes with alternative TSSs than for those having single TSSs, indicating that the presence of condition associations for more than one core promoter is a common phenomenon across all conditions. Because we assigned conditions to individual TSSs, it was possible for the 1,225 genes

with alternative TSSs to have more than one association. We thus divided genes with alternative TSSs into two groups: genes whose TSSs had different condition associations, if at least one TSS had at least one different association from the gene's remaining TSSs, and genes with the same condition associations for all of the alternative initiation sites. In our dataset, 392 (32%) genes with alternative TSSs had the same condition association, and over two times that number of genes with alternative TSSs, 833 (68%), had different condition associations. The association of *tramtrack's* TSSs to the embryo exemplifies typical patterns seen for the set of 392 genes with alternative TSSs having the same condition associations.

4.3.1.3 Quality Assessment of Spatiotemporal Associations

The two main sources of library bias that can affect the determination of the condition specificity of the TSSs by applying Shannon entropy to the ESTs are library normalization and library size. While we recognize that a slight normalization bias may remain in each of the individual libraries, we do not believe that the condition specificity assignments were significantly impacted by these different biases because multiple independent libraries were used to generate the collections of ESTs for each of the eight conditions. It is highly unlikely that the exact same normalization bias would be present in all of the independent libraries. The second source of bias that may affect TSS condition specificity arises from using libraries of different sizes. The repertoire of ESTs

used in this work is the largest amount of publicly available transcript data for *Drosophila* to date, but certainly does not saturate the transcriptome. As a consequence, differences in the numbers of ESTs from each cDNA library may affect the condition associations made using Shannon entropy. In an effort to minimize this bias, we removed alternative TSSs with very few ESTs because they had a higher probability of being affected by library size bias. We also penalized entropy values according to the sampling depth of ESTs in their corresponding libraries. In spite of these efforts, alternative TSSs may have condition associations simply due to the low resolution of the available data. For instance, alternative TSSs may have embryo specific associations using the EST library information, but be utilized at different temporal stages of embryogenesis. We confirmed this phenomenon to some extent for the group of TSSs active during embryogenesis by using available expression data from whole-genome tiling arrays. However, higher resolution of data is needed across all body parts and time periods to gain deeper insight into the precise spatiotemporal utilization of TSSs.

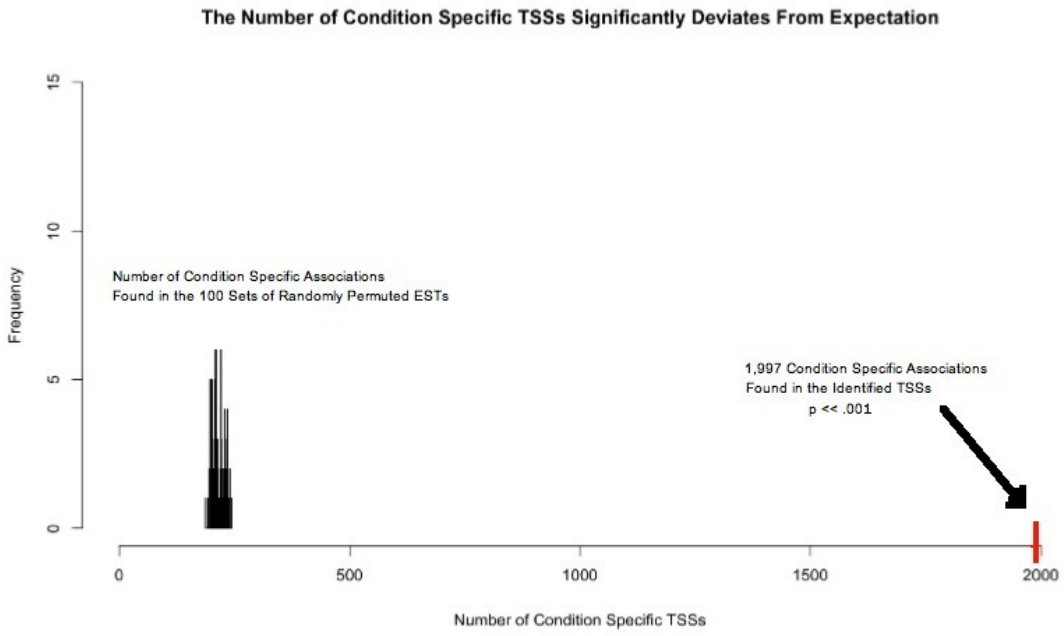


Figure 36: Condition Associations For Random Permutations of Labels

Condition assignments were repeated on 100 sets of random permutations of the 72,535 condition labels across the 5,665 (sub-) clusters. The total number of sites with specific condition associations was summed for each permutation. Across all 100 sets of permutations, the number of condition specific sites ranged from 180 to 250. The 1,997 condition specific TSSs in the identified set significantly deviated from this distribution ($p << .001$).

The number of 1,997 condition-specific TSSs that we identified was significantly higher than random permutations ($p << .001$) (see Materials and Methods, Figure 36). The mean number of condition specific sites in the set of 100 random permutations was 215, leading to a false positive estimate of 10.8%. In turn, there were significantly fewer sites with mixed associations and condition-supported TSSs ($p << .001$ for both groups). We empirically estimated the number and rate of false positives for each condition

individually (Table 2). These results indicate that any bias resulting from differences in library sizes is minimal, even for the smallest libraries. The numbers of condition specific TSSs identified greatly deviate from expectation, and are thus, likely to reflect true patterns of transcription initiation.

Table 2: False Positive Estimates of TSS Assignments by Condition

To assess the validity of the TSS condition assignments, we performed 100 random permutations of condition labels from the (sub-)clusters and evaluated their associations using the same methodology as for the identified TSSs. The numbers of false positives (column 3) were empirically estimated as the mean number of sites having a specific association to each condition (column 1) across all 100 random permutations. The false positive rate (column 4) was calculated by dividing the number of false positives by the number of identified TSSs that were observed to have the condition association (column 2).

Specific TSS Association	No. of Identified TSSs	Estimated No. False Positives	False Positive Rate
Embryo	1134	152	13.40%
Larva Pupa	27	0	0%
Head	522	63	12.10%
Ovary	9	0	0%
Testes	247	0	0%
Schneider Cells	36	1	2.80%
Mbn2 Cells	12	0	0%
Fat Body	10	0	0%

When we evaluated the significance of the gene associations, the number of genes with alternative TSSs having the same condition associations was significantly higher than random expectation ($p << .001$, see Materials and Methods). This implies

that alternative TSSs of the same gene have a higher than expected probability of being utilized in the same condition than in different conditions. However, with additional conditions and ESTs, a larger percentage of alternative TSSs with different associations may result.

We also performed a Gene Ontology (GO) analysis to assess whether genes with alternative TSSs had enrichments for specific functional categories based on their distinct conditions (Ashburner et al. 2000). GO assessments were made for each of the eight distinct conditions (Ashburner et al. 2000), and whole genes with more than one TSS or TSS cluster group were designated as having alternative TSSs. The foreground set consisted of genes with alternative TSSs associated with a specific condition. The background set included genes with associations to that condition plus all genes with a mixed condition association. We evaluated enriched terms on levels 2 and 3 of the GO hierarchy. The overrepresentation of terms in the foreground set when compared to the background set was measured using a hypergeometric test. To control for multiple comparisons, we applied a false discovery rate (FDR) threshold of 10% (Klipper-Aurbach et al. 1995). We also excluded terms with less than or equal to five expected occurrences. This showed enriched terms broadly reflecting functions required for the proper development of the embryo, suggesting that genes with alternative TSSs are significantly active during embryo-specific processes (Table 3).

Table 3: GO Enrichments

The table lists all significant GO categories at a false discovery rate cutoff of 0.1 that are present in more than five genes, for genes with alternative TSSs associated with specific conditions.

Condition	P value	FDR	No. Observed	No. Expected	GO Term
Embryo Biological Process:					
Level 2:					
	0.000221	0.002040	35	19	system process
	0.000572	0.004081	26	14	behavior
	0.002412	0.006122	54	37	cellular devel. process
	0.002576	0.008163	24	14	cell motility
Molecular Function:					
Level 3:					
	4.29E+06	0.002380	29	12	cytoskel. protein binding
Cell Component:					
Level 2:					
	0.000626	0.004761	60	40	membrane of cell
Level 3:					
	0.000217	0.001923	27	14	plasma membrane
	0.000386	0.003846	22	11	plasma membrane part
Testes Cell Component:					
Level 3:					
	0.000241	0.003448	16	8	cytoplasm

As discussed in the Background, a number of well-known genes involved in development are known to have developmentally regulated alternative TSSs associations. In *Drosphila*, there have been overall few well-studied examples of genes having differentially utilized promoters, compared to human and mouse (Landry, Mager, and Wilhelm 2003). This is especially true for the differential regulation of TSSs during embryogenesis, as the transition to zygotic transcription is known to cause the degradation of maternal mRNAs (De Renzis et al. 2007; Schier 2007). While some of the well-known genes have too few ESTs to call their TSSs, we collected a list of 10 genes with known utilization during embryonic development, as verified by *in situ* and published sources (Table 4). The patterns of embryonic utilization for these genes in our dataset confirm previous observations and further suggest that the promoters of TSSs for these genes are developmentally regulated.

Table 4: Embryo Associations Confirm Utilization Patterns of Known Genes

We compared the embryonic utilization patterns previously observed for known genes to those identified using EST and Affymetrix tiling array data. Analysis of genes with at least one TSS having an EST embryo association (column 3), and promoter utilization in at least one tiling array time period (column 4) agree with previously reported expression patterns from in situ images (column 5) (Tomancak et al. 2002), and published reports (column 6).

Gene ID	Name	EST Embryo Assoc	Tiling Array	BDGP in situ	Developmental Regulation
CG10334	spi	X	X	X	Rutledge et al, Genes Dev 1992
CG1856	ttk	X	X	X	Read et al, Mech Dev, 1992
CG2671	l(2)gl	X	X	Klaembt et al, EMBO, 1986	Mechler et al, EMBO, 1985
CG31243	cpo	X	X	X	Bellen et al, Genes Dev, 1992
CG3725	Ca-P60A	X	X	X	Varadi A et al, FEBS Lett, 1989
CG4898	Tm1	X	X	Hales et al, Dev Biol, 1994	Hales et al, Dev Biol, 1994
CG8989	His3.3B	X	X	Feng et al, Genome, 2005	Akhmanova et al, Genome, 1995
CG9075	eIF-4a		X	Hernandez et al, Proteomics, 2004	Dorn et al, Mol Gen Genet, 1993
CG9261	nrv2	X	X	X	Xu et al, Gene 1999
CG9553	chic	X	X	Cooley et al, Cell, 1992	Cooley et al, Cell, 1992

Associations were also confirmed for TSS associations to the other conditions. For the gene *Calmodulin* (CG8472), we identified two TSSs separated by 21bp. The upstream TSS had a mixed association, while the downstream promoter had LPS induced mbn2 support. Utilization of the downstream promoter agrees with the known activation of *Calmodulin* as part of the LPS induced mbn2 immune response (Loseva and

Engstrom 2004). For the gene CG11151, we also identified two TSSs separated by 100bp; the upstream TSS had a mixed association, and the downstream TSS was supported by the larval fat body, confirming reported expression of the gene in the lipid subproteome (Beller et al. 2006). Condition associations may also provide new information about genes with limited or no molecular, biological, or cellular functional evidence in Flybase. As examples, the gene CG10510 had one TSS with a specific association to the testes, and the TSS for CG1814 showed support for the head condition.

4.3.2 Differences in the Temporal Utilization of Alternative Promoters During Embryogenesis

4.3.2.1 Patterns of Alternative TSSs Are Distinct

While we observed a significant enrichment of alternative TSS associations to the same conditions, EST libraries are too broad to distinguish differences in the precise timing of a promoter's temporal utilization. To examine initiation events at higher resolution, we used available Affymetrix whole-genome tiling arrays of *D. melanogaster* embryonic expression. The data was a natural fit to our analysis because expression of genes was monitored at 12 time points during the first 24 hours of the developing *D. melanogaster* embryo, each covering a 2 hour period (Manak et al. 2006). Embryogenesis has been well studied in *Drosophila*, and the morphological changes that occur have been examined in depth. Transcriptional control in early embryogenesis, involving well

known transcription factors, such as *kruppel* and *eve*, is an important model system for gene regulation in development that has greatly advanced the field (Alberts et al. 2002).

To examine initiation events at a high resolution, we evaluated activity of 2,765 genes with one peaked promoter, 685 genes with one broad promoter, and 540 genes with a combination of promoter types using available Affymetrix whole-genome tiling arrays of *D. melanogaster* embryonic expression (Manak et al. 2006) (see Materials and Methods, Appendix D). By pooling all promoters together, we saw 58.7% transcribed in at least one of the 12 embryonic time points. The largest number of promoters (1,640 and 1,455, respectively) was utilized at time points 1 and 2, compared to any other developmental period (Figure 37). These results agreed with previous analyses of the tiling data which focused on whole transcripts (Manak et al. 2006). At this early stage in development, the majority of promoters are expected to correspond to maternal utilization. There was a decrease in the number of promoters utilized at point 3, followed by a second maximum of ~1,300 promoters utilized at time points 5 and 6. This corresponded to the decrease in maternally inherited transcripts and the initiation of zygotic transcription. After time point 6, the number of promoters utilized continued to decrease, with a third weaker maximum at period 11, signifying late zygotic transcription. The onsest and duration of expression patterns across the 12 time points in three transcription cycles: 1-2, 4-8, 9-12 was previously (Manak et al. 2006). It suggests the existence of periods during which transcription factor binding and/or RNA pol II

activity changes simultaneously for large numbers of genes during embryogenesis. Further statistical analysis is needed to rigorously evaluate the significance of this trend.

There was a tendency for more frequent patterns to show expression at contiguous time points and to start and/or stop at cycle boundaries. The most frequent patterns for all promoters (peaked and broad) were "all off", i.e. no utilization during any period (41%), and "all on", i.e. expression for the entire 24-hour duration of embryogenesis (5.8%; 272 TSSs). This provides support that these patterns are not artifacts of the tiling array processing, but rather, true promoter utilization over longer time intervals. Some patterns showed expression for more than one contiguous time point. This may result from not detecting transcription for cases in which the expression level fell just below the determined threshold. The pattern may also reflect the utilization of more than one promoter that we were unable to differentiate due to the resolution of the tiling array. For instance, a common pattern had activity at four time points: T,T,N,N,T,T,N,N,N,N,N,N (T=transcribed; N=non-transcribed). For overlapping promoters in a TSS cluster group, this pattern may be the result of the combination of utilization of one promoter for a maternally inherited transcript during periods 1 and 2, and another promoter utilized for early zygotic transcription during periods 5 and 6. Temporal patterns observed for more than five promoters are listed in Appendix E.

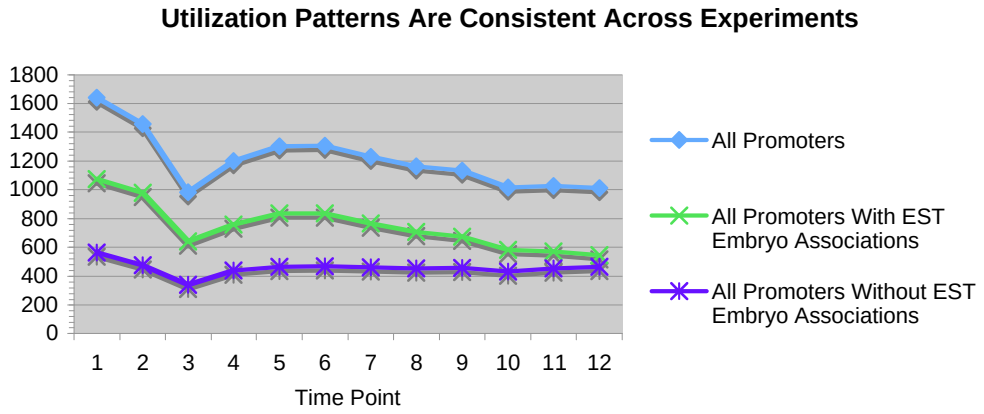


Figure 37: Consistent Trend of Embryonic Utilization as Measured by Affymetrix Tiling Arrays Across EST and Tiling Experiments

Median differences in tiling array fluorescence intensities were used to detect transcription at 4,664 peaked and broad promoters. The total number of transcribed sites was plotted for each of the 12 time points, corresponding to two-hour increments during embryogenesis. The promoters were separated into two groups at each time point: those with embryo EST associations and those without.

When comparing promoter utilization to the EST associations, there were 2,558 sites with embryo EST associations, and 2,106 without. On average, there were 1.6 times more sites with embryo EST associations detected as transcribed by the tiling array at each time point than those with non-embryo EST associations (head specific, testes supported, etc) (Figure 37). When normalized by the total number of sites with embryo and non-embryo EST associations, this is a statistically significant trend ($p < .01$, paired t-test). As both experiments were performed on the embryo, this confirms that measurements of transcript expression are consistent across data types. There were 68.4% of sites with embryo associations and 46.9% of sites with non-embryo EST

associations whose promoters were utilized in at least one of the 12 2 hour periods, as determined by the tiling array. The latter group contained sites with "mixed" associations, many of which are from the Exelexis EK library, which was generated from embryos, imaginal discs, and adult heads (Stapleton et al. 2002). Thus, active initiation sites with non-embryo EST associations most likely included some that are, in fact, active in embryonic transcription programs. The difference between the utilization of promoters with tiling support and those with and without embryo EST associations is greatest during the first two time points, decreases throughout the remaining time points, and disappears at time point 12 (Figure 37).

Temporal biases of transcriptional activity were seen in the tiling array when the total number of promoters was divided into peaked and broad. After normalization by the total number of promoters in each set, a statistically significant higher fraction of broad promoters were utilized than peaked promoters in the tiling array ($p << .01$, Figure 38**,** see Material and Methods). The difference was greatest in the first and second 2-hour periods, and reached an additional maximum at time points 5 and 11. While it continued to decrease after time point 5, the difference remained through time point 12. Overall, 56.6% of peaked promoters were transcribed in at least one of the 2-hour periods and, 67.8%, or 11.2% more, broad promoters were transcribed in at least one period.

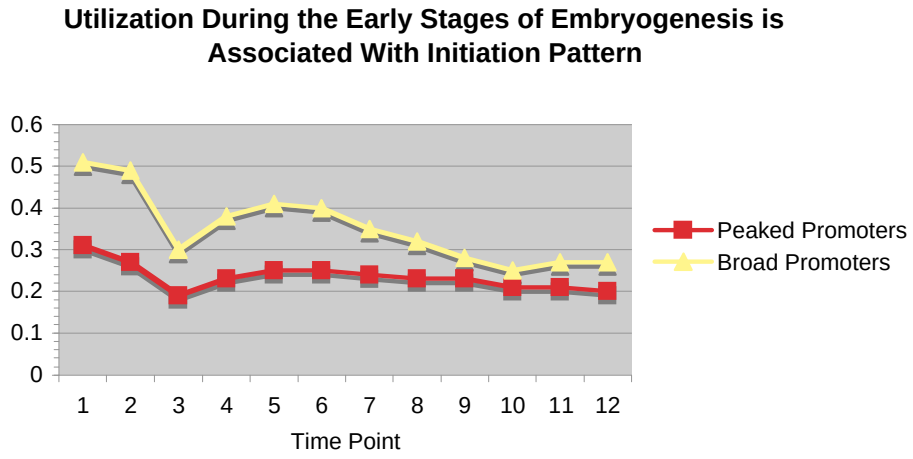


Figure 38: Promoter Types Are Correlated With Timing of Utilization

The set of all promoters was divided into 3,788 peaked and 876 broad. At every time point, the fractions of transcribed peaked and broad promoters were found by dividing the number of transcribed promoters in each group by the total number of peaked and broad promoters, respectively.

The pattern that broad promoters were more transcriptionally active during embryogenesis than peaked promoters was separately mirrored using the EST associations alone, without the tiling array data ($p << .01$, see Materials and Methods). Here, initiation sites were deemed to have an embryo EST association if an individual TSS, or at least one of the TSSs in a TSS cluster group had the association, resulting in 50.3% of TSSs and 74.3% of the TSS cluster groups having embryo-specific or embryo supported associations. When comparing the condition associations of both promoter types across EST and tiling array experiments, we saw consistency in embryonic

utilization of promoters. Overall, 1,682 peaked and 288 broad promoters showed no utilization during any of the 12 developmental time points.

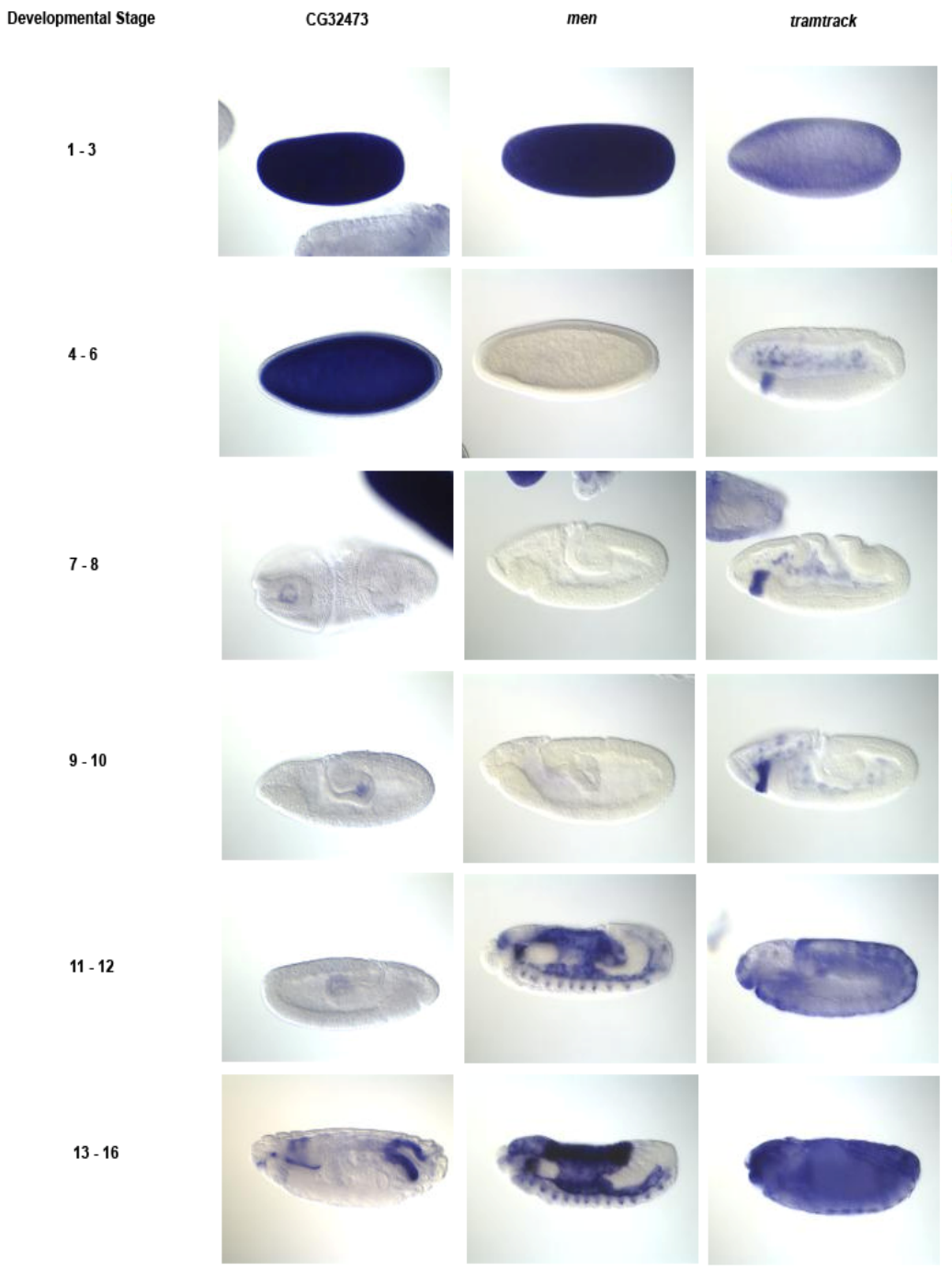
4.3.2.2 Genes With Alternative TSSs

Next, we explored the profiles of genes with alternative promoters in greater depth. In this analysis, we excluded broad promoters from the set of 540 genes with alternative TSSs separated by at least 100bp, on account of their lack of precise individual TSS resolution, and divided the remaining 407 genes into four categories. The first category consisted of 143 genes (35%) with no expression from any peaked promoters at any time point. The second category comprised 170 genes (42%) with exactly one alternative promoter active during embryogenesis. In this group, 75 genes showed expression at time point 1 and their promoters were thus maternally utilized. In the third category, there were 20 genes (5%) with more than one, but less than all alternative peaked promoters having utilization during embryogenesis. The remaining 74 genes (18%) in the fourth category had all alternative peaked promoters utilized at some time during embryogenesis.

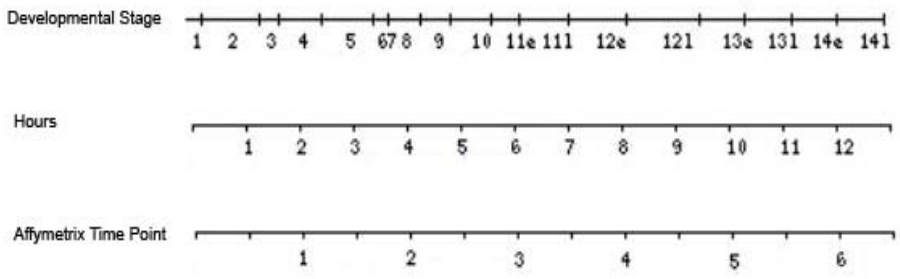
For the 74 genes in the fourth group, we examined the onset of utilization, as defined by the first time point in which utilization lasted at least 4 hours, or 2 periods. This removed isolated and thus potentially erroneous calls. There were 30 genes with the same onset time across alternative peaked promoters, albeit different durations of

utilization. The temporal utilization of the 44 genes with different onset across alternative peaked promoters was typically a combination of both maternal and zygotic utilization. For two candidate genes in particular, CG10120 (*men*), and CG32473, different peaked promoters corresponded to completely non-overlapping periods of activity. Available RNA *in situ* images (Tomancak et al. 2002) beautifully illustrated that the activity of distinct alternative promoters is associated with different spatiotemporal expression patterns (Figure 39). This switch in maternal versus zygotic promoter utilization mirrors the transcription of the well-studied gene *hunchback*, for which our dataset unfortunately did not contain enough ESTs to call TSSs. This analysis shows that dynamic properties of alternative promoter activity, such as onset and duration, are needed to properly characterize the regulation of transcription initiation during embryogenesis.

A


Developmental Stage
CG32473
men
tramtrack
1 - 3
4 - 6
7 - 8
9 - 10
11 - 12
13 - 16

B



C

Gene	TSS	Affymetrix Time Point (T = Transcribed, N = Not Transcribed)											
		1	2	3	4	5	6	7	8	9	10	11	12
CG32473	9128375	N	N	N	N	N	N	N	N	T	T	T	T
	9130491	T	T	N	N	N	N	N	N	N	N	N	N
	9135107	T	T	N	N	N	N	N	N	N	N	N	N
men	8545514	N	N	N	T	T	T	T	T	T	T	T	T
	8548153	T	T	N	N	N	N	N	N	N	N	N	N
tramtrack	27539771	N	N	T	N	T	T	N	N	N	N	T	N
	27550733	T	T	T	T	T	T	T	T	T	N	N	N
	27551504	N	N	T	T	T	T	N	N	N	N	N	N

Figure 39: Differences in the Temporal Activity of Alternative TSSs Correspond to Distinct Patterns of Gene Expression

(A) In situ Expression Patterns of Genes with Alternative TSSs In situ images showing the spatiotemporal expression of the CG32473, CG10120 (men), and CG1856 (ttk) genes during development (Tomancak et al. 2002). **(B) Correspondence Between Time Period and Developmental Stage** As reference, the timing of developmental stages of the *Drosophila* embryo is matched to a timeline of one-hour intervals and the Affymetrix 2-hour increment time course. **(C) Utilization Patterns as Measured by the Tiling Array** The TSSs identified from the most frequent 5′ EST ends are listed for each gene. The patterns of peaked promoter utilization detected on the tiling array are noted according to the 12 time points measured during embryonic development. Tiling array data showed that the peaked promoter of TSS #1 was utilized at time points 3, 5, 6 and 11 (hours 4-6, 8-12, and 20-22), TSS #2 at 1-9 (hours 0-18), and TSS #3 was used at time points 3-6 (hours 4-12). While the pattern of utilization of the promoter of TSS #1 flipped at time points 4 and 11, the patterns for both TSS #2 and #3 were contiguous. TSS#2 is maternally inherited and the utilization of its promoter extends through early zygotic stages, while the utilization of the others starts after four hours and is active for a shorter time. Notably, the peaked promoter of TSS#2 was the only one without an INR motif.

There were 141 genes whose alternative TSSs were more than 100bp apart and had the same EST condition associations. For 26 of these genes, no promoters were active at any time point. Only five genes had promoters with the same activity pattern, and all five showed activity across all 12 time points. The remaining 110 (78%) genes had different temporal patterns of utilization across alternative promoters. This clearly demonstrates that while associations may be the same across larger global conditions, such as those corresponding to the EST libraries, data on a more precise scale may reveal differences in initiation patterns.

All three peaked promoters of the *tramtrack* gene were separated by at least 100bp and each had an EST association to the embryo. Typical of the set of genes with

the same EST conditions, temporal analysis of the alternative promoters revealed different patterns of utilization. Figure 39 shows the tiling array utilization and *in situ* staining of the complex patterns of gene expression observed for *ttk* during each stage of embryogenesis. While further experimental verification is needed to decipher the association between the spatiotemporal patterns and the utilization of each of *ttk*'s alternative promoters, RNA *in situ* images show the existence of distinct expression patterns at different stages that are consistent with the usage of alternative promoters (Tomancak et al. 2002).

4.3.13 Quality Assessment of Temporal Promoter Associations by Tiling Arrays

The fluorescence intensities of random sites were used to estimate the rate of false positives. The number of 100bp sequences surrounding random sites that were deemed transcribed ranged from 92 to 162 across all 12 periods, resulting in a low expected false positive rate of .02 to .035 (Table 5). This agrees with the .02-.046 rate of transcription previously observed in intergenic regions across the 12 periods (Manak et al. 2006). The true false positive rate may be lower than this approximation as some random intergenic locations may correspond to unidentified exons, or the expression of other genomic elements.

Table 5: False Positive Estimates for Embronic Temporal Promoter Assignments

We evaluated the expected number of false positive temporal expression assignments for the set of promoters of 4,664 identified TSSs across 12 developmental periods (column 1) corresponding to 2 hour increments during embryogenesis (column 2). We chose 4,664 random intergenic sites and found the difference in median fluorescence intensities of neighboring tiles for each of the 12 time points. The differences in fluorescence intensities were compared to the difference thresholds (column 3) used to classify the set of 4,664 promoters. Random intergenic sites with fluorescence intensity differences above the threshold were counted as false positives. For each time point, the total number of false positives (column 4) was divided by the total number of random intergenic sites to approximate the rate of false positives (column 5).

Time Point	Hours	Difference Threshold	Number of False Positives	False Positive Rate
1	0-2	36.5	103	0.022
2	2-4	23.5	94	0.020
3	4-6	21.0	158	0.034
4	6-8	23.5	121	0.026
5	8-10	29.5	122	0.026
6	10-12	35.0	126	0.027
7	12-14	24.5	113	0.024
8	14-16	32.0	162	0.035
9	16-18	28.0	113	0.024
10	18-20	21.0	149	0.032
11	20-22	23.0	92	0.020
12	22-24	18.5	93	0.020

In comparison, Manak et.al (Manak et al. 2006) approached promoter utilization for TSSs from the same tiling data computationally by first clustering transfrags and manually curating 5′ start sites. Two disadvantages of this approach are lower TSS resolution in identifying TSS locations, and lower accuracy of clustering assignments. The average and median lengths of the transfrags used were 328 and 197bp,

respectively, which corresponds to at least two non-overlapping core promoters, with additional sequence in between. Their clustering strategy had a reported accuracy of 65-77% when the number of correctly identified new first exons of known genes from the tile fluorescence intensities were evaluated by sequencing.

4.3.3 Core Promoters of Maternally Inherited and Zygotically Active TSSs Have Characteristic Profiles of Sequence Elements

The presence of the two types of core promoters defined by different initiation patterns in *Drosophila* and vertebrates suggests that each may have a functional importance. To determine potential associations to specific conditions, we first compared the motif composition of 370 peaked promoters with head specific TSS EST associations, and 765 peaked promoters with embryo specific TSS EST associations (see Materials and Methods Chapter 3). Broad promoters were excluded from this analysis because each of the TSSs in the TSS cluster groups could have a different EST association, and due to the weak spatial biases of some motifs, distinct assignments of elements to conditions would be difficult for overlapping core promoters. In both promoter sets, the TATA, INR, DPE, and MTE had higher probabilities of occurring in their preferred windows than the Ohler 1, DRE, Ohler 6, and Ohler 7 elements. This mirrored the higher enrichment signal for the location specific motifs. However, while we saw small differences between motif frequencies in the embryo and head specific promoters, no clear trends for condition-enriched motifs were observed, and therefore,

these associations were not investigated further. This most likely resulted from the low resolution of these conditions, as both "head" and "embryo" encompass numerous tissues across various developmental stages.

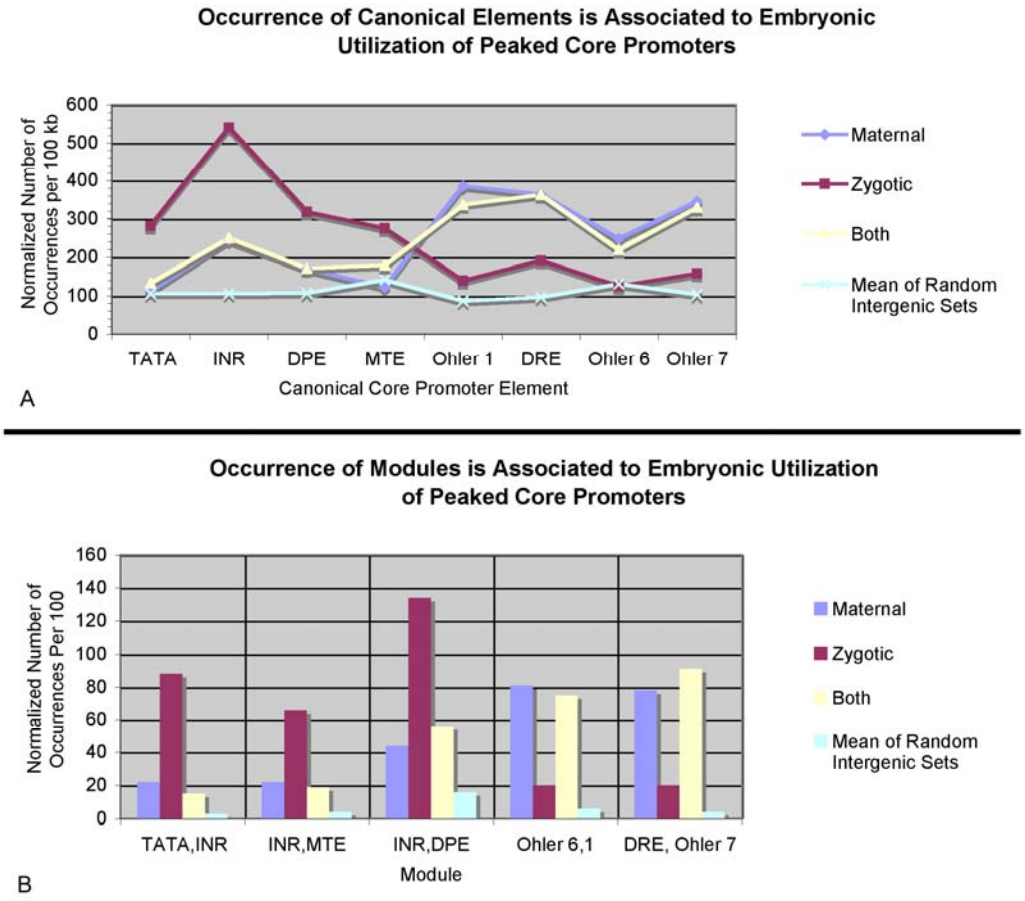


Figure 40: Elements in Peaked Promoters Are Associated to Embryonic Utilization

(A) Maternal and Zygotic Activity of Peaked Promoters Corresponds to Differences in Element Occurrences The presence of eight sequence elements was evaluated in peaked core promoters of TSSs using PATSER. Core promoters were segregated into three groups based on their pattern of utilization (maternal, zygotic, both). Those showing no expression during the time course were excluded from this analysis. The normalized means of motif matches in three random intergenic sets are shown. **(B) Regulatory Modules Also Segregate By**

Condition For Peaked Promoters The numbers of occurrences of motif modules were evaluated in each of the three groups of peaked core promoters (maternal, zygotic, both) by counting the numbers of pairs of matches positioned in the designated order, with respect to the orientation of transcription.

We therefore examined the presence of sequence elements in the more precisely defined conditions that the tiling expression time course data allowed for, and analyzed 319 maternally inherited, 766 zygotically utilized, and 1,021 mixed maternally and zygotically active peaked promoters (see Materials and Methods). We performed a concurrent analysis on 97 maternally inherited, 99 zygotically utilized, and 392 mixed broad promoters, to ensure that any identified associations of promoter elements to embryonic time points were consistent for different initiation patterns. The set of zygotically utilized peaked promoters showed a clear enrichment in the elements with strong positional bias - the TATA, INR, DPE, and MTE - and the maternally utilized sites had higher frequencies of the less location biased elements (Ohler 1, DRE, Ohler 6, and Ohler 7; see Figure 40A). While smaller differences in the frequencies of the elements were observed in the broad promoters overall, the same pattern of motif matches in the maternal versus zygotic conditions was found (see Figure 41A). The association of the DRE, Ohler 6, and Ohler 7 motifs to maternal utilization was supported by a previous motif analysis that evaluated the significance of ImaGO terms in the *Drosophila in situ* hybridization database (Down et al. 2007). As this division in motif usage for maternal vs. zygotic transcription was observed for both initiation patterns, it indicated that the

repertoire of elements in the core promoters is determined by the different conditions. To test the relationship between initiation pattern and condition, we summed the normalized frequencies of the location specific motifs (TATA, INR, DPE, and MTE) and non-location bias motifs (Ohler 1, DRE, Ohler 6, Ohler 7) in peaked promoters with maternal (resp. zygotic) utilization, and in broad promoters with maternal (resp. zygotic) utilization, and performed a χ^2 test on both 2 X 2 contingency tables. In both χ^2 tests, the null hypothesis that initiation patterns and temporal conditions are independent of each other was rejected at ($\alpha = .05$), indicating that maternal vs. zygotic activity of core promoters and their initiation patterns are related to each other.

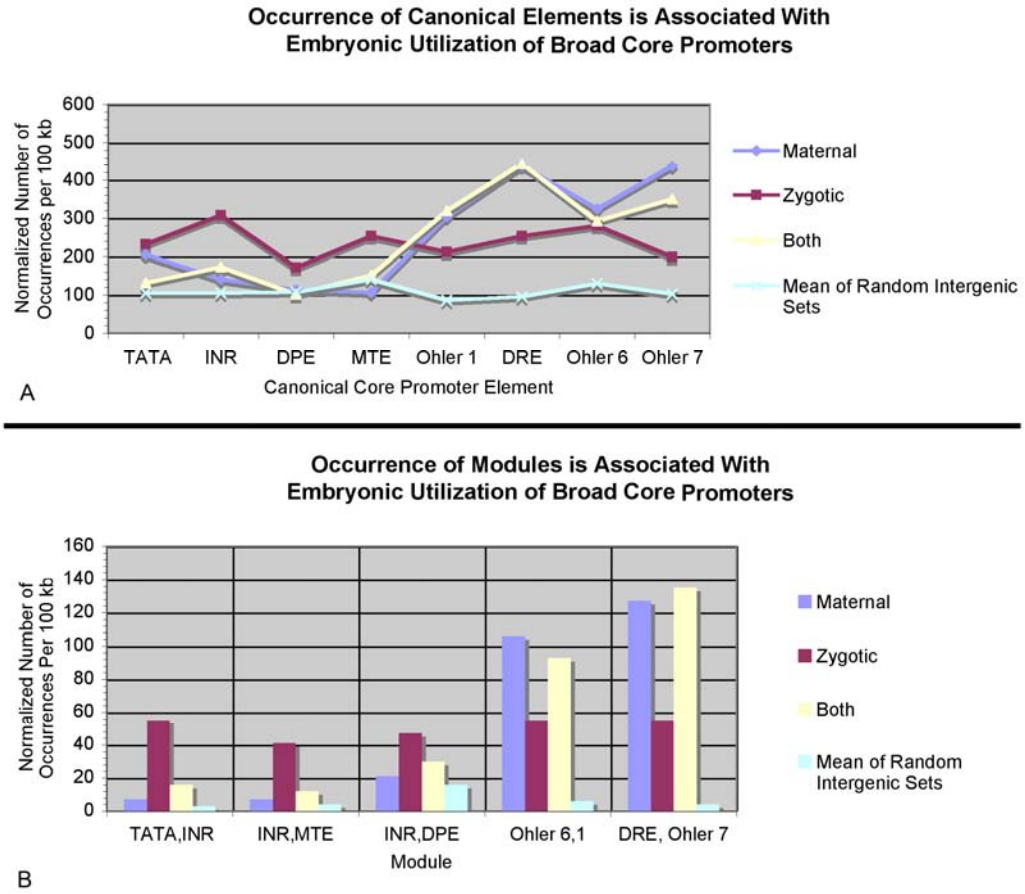


Figure 41: Motif Elements in Broad Promoters Maintain Pattern of Embryonic Utilization

(A) Maternal and Zygotic Activity of Broad Promoters Corresponds to Differences in Element Occurrences Broad core promoters were segregated into three groups based on their pattern of utilization (maternal, zygotic, both) and normalized motif occurrences were found using PATSER, akin to that of peaked promoters. **(B) Regulatory Modules Also Segregate By Condition** For Broad Promoters The numbers of occurrences of motif modules were evaluated in each of the three groups of broad core promoters (maternal, zygotic, both) akin to that of peaked promoters and are reported here.

For the set of peaked promoters with both maternal and zygotic utilization, slight fluctuations in element frequencies of the TATA, MTE, Ohler 1, and Ohler 7 were seen,

however, the differences were not large enough to alter their overall pattern that mirrored that of maternal utilization (see Figure 40A). There were three times more sites with utilization during both phases of embryogenesis containing the Ohler 1, DRE, Ohler 6, and Ohler 7 motifs, than those having maternal utilization alone. When the sequence elements were restricted to their preferred windows in the peaked core promoters (see Materials and Methods Chapter 3), the same trends of maternal versus zygotic element preference were observed (see Figure 42). Similar results were seen for broad promoters (see Figure 41A). This relationship can be expected, as promoters with both patterns of utilization could in fact have resulted from the use of maternal promoters whose transcripts were not yet degraded within the cell. In both promoter types, the TATA and INR had the highest motif frequencies, and the Ohler 6 and Ohler 7 had the lowest. This confirms the importance of location for the TATA and INR, and the presence of the Ohler 6 and Ohler 7 motifs throughout promoters utilized both maternally and zygotically. When compared to the numbers of occurrences in the random intergenic sets, the frequencies of the most common motifs were much higher overall in the promoters, although some of the less common motifs were in the range of frequencies observed for the random sites. This shows that when not in proper context, occurrences of the sequence elements are not as meaningful.

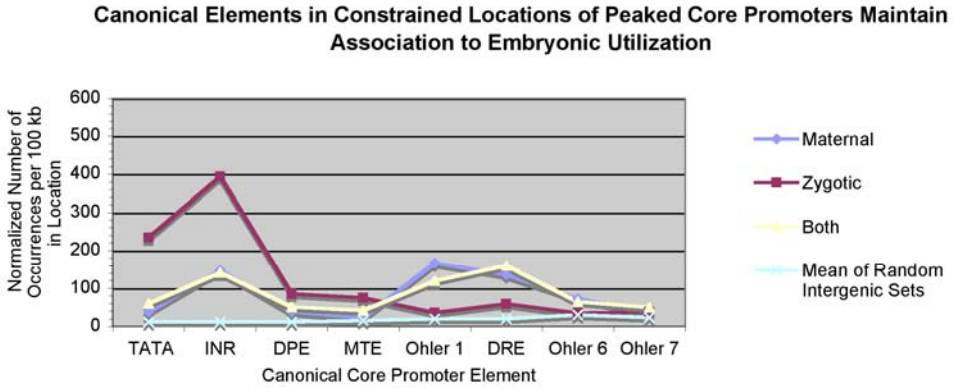


Figure 42: Sequence Elements in Preferred Windows of Peaked Promoters Preserve Associations to Embryonic Utilization

The set of peaked core promoters was divided into three groups according to the their pattern of embryonic utilization (maternal, zygotic, or both). The numbers of canonical elements located in the preferred windows of core promoters were counted and normalized for each group as shown.

Akin to individual motif analysis, the occurrences of the TATA/INR, INR/MTE, INR/DPE, Ohler 6/1, and Ohler 7/DRE modules were evaluated separately for maternal and zygotic utilization (see Materials and Methods). The results showed that the TATA/INR, INR/MTE, and INR/DPE had higher frequencies in the zygotically transcribed peaked promoters, and the Ohler 6/1 and Ohler 7/DRE had higher frequencies in the maternally utilized peaked promoters. Similarly, the numbers for peaked promoters with both maternal and zygotic transcription initiation agreed with the maternally utilized module frequencies (see Figure 40B). The same trends were observed for broad promoters (see Figure 41B). In summary, these findings therefore complement the associations of initiation patterns to motifs, and propose that specific

core promoter elements are more frequently utilized during different stages of development.

4.4 Discussion

Alternative TSSs that are active under different spatiotemporal conditions have been reported for several individual fly genes (Corbin and Maniatis 1989; Margolis et al. 1994). Our analysis here established distinct spatiotemporal utilization of alternative TSSs as a common feature in *D. melanogaster*. The results strongly indicate that usage of many alternative TSSs is condition-dependent. In humans, previous work has shown that the aberrant use of alternative promoters is associated to various diseases, such as cancer (Davuluri et al. 2008). Genomic similarities that can be observed in the usage of alternative TSSs under different spatiotemporal conditions in both humans and *Drosophila* may provide insights into the mechanisms governing disease.

Alternative TSSs may also be utilized under the same broad EST-derived conditions. In fact, there was a higher than expected number of genes with alternative TSSs with the same condition association. Alternative TSSs with the same condition associations may result from a series of point mutations, or be created anew through promoter sequence duplication (Frith et al. 2006). Alternative TSSs may also be associated with the same condition simply due to the low resolution of the available data. For instance, alternative TSSs may be deemed embryo specific using the EST

library information, but be utilized at different temporal stages of embryogenesis. We confirmed this phenomenon to some extent for the group of genes/TSSs active during embryogenesis by using available expression data from whole-genome tiling arrays.

Recent work has shown that core transcriptional complexes can be remodeled in specific cell types in both mammals and flies (Deato and Tjian 2007; Metcalf and Wassarman 2007). As consequence, some possibly yet unknown core promoter elements may have specific associations to the spatiotemporal conditions analyzed here. In our study, we showed that peaked *D.melanogaster* promoters are utilized zygotically, confirming previous findings that the promoters of genes with the INR and DPE are associated to developmental regulation and that the TATA is overrepresented in terminally differentiated tissues, such as the cuticle, and endocrine glands (Engstrom et al. 2007; FitzGerald et al. 2006). Developmentally regulated genes were later shown to be associated to the stalling of the RNA pol II machinery in *D.melanogaster* (Zeitlinger et al. 2007). In agreement with this finding, vertebrate peaked promoters are known to have an association to more tightly regulated transcripts (Carninci et al. 2006). These associations between motifs may reflect larger scale organization of the transcriptional machinery. A circuit involving the TATA binding protein (TBP), Mot1, and NC2 was shown to exist that controls the regulation of DPE-dependent versus TATA-dependent transcription (Hsu et al. 2008). This suggests that a larger network regulates the transcriptional balance between functional classes of core promoters. As this analysis

characterized individual sites of transcription initiation, and previous studies evaluated associations using whole genes in *D. melanogaster*, the functional associations of peaked promoters to developmental regulation and terminally differentiated tissues should be explored in greater depth.

By showing that broad promoters are maternally utilized in *D. melanogaster*, this work supported previous studies showing that core promoter motifs without a location enrichment are utilized in the embryo, are associated to housekeeping functions, such as DNA repair and translation, and the proteins necessary to perform them, such as the components of the RNA pol II, and mitochondrial proteins (Engstrom et al. 2007; FitzGerald et al. 2006). Housekeeping genes with ubiquitous expression are associated with actively transcribing pol II in *D.melanogaster* (Zeitlinger et al. 2007), and with broad patterns of initiation in vertebrates (Carninci et al. 2006). The association of broad promoters to maternal utilization also suggests the hypothesis that larger regions of the DNA may be accessible at these locations. The localization of nucleosomes or specific chromatin marks may affect the accessibility of the DNA under specific conditions and locations, and explain the presence of these initiation patterns (Mavrich et al. 2008; Mito, Henikoff, and Henikoff 2005). In support of this hypothesis, a previous study suggested that the promoters recognized by TRF2 up-regulate genes are required for specific developmental pathways and may be involved in chromatin organization in mammalian gonads (Isogai et al. 2007). In addition, the study showed that the positional bias motifs

were associated to the recognition of the core histones H2A/B, H3/H4 gene promoters by TBP and that the motifs with a weak positional bias were associated to the transcription of the linker *Histone H1* gene promoter (Isogai et al. 2007). An alternative model is that different TAFs or other proteins interacting with the RNA pol II may be responsible for maternally initiating the transcription of genes at broad regions throughout the promoter.

Additional verification is needed for the spatiotemporal utilization of peaked and broad promoters. A possible experimental validation of specific expression patterns linked to alternative promoters includes RNA *in situ* hybridization during different stages of fly embryogenesis (Tomancak et al. 2002; Tomancak et al. 2007). *In situ* images are able to capture spatial gene expression patterns at a much higher resolution than ESTs and microarrays. Our study provided promising candidates to design isoform specific probes, which would link differences in the spatial and temporal expression of transcripts for the same gene to different promoters. However, *in situ* hybridization requires experimentation on individual TSSs, and additional data is needed across all body parts and time periods to gain deeper insight into the precise spatiotemporal utilization of TSSs. As high throughput 5′ tags during the 0-24hr of embryonic development have become available in *D.melanogaster* concurrently during the course of this study (Ahsan et al. 2009), the exploration of TSS utilization under different spatiotemporal conditions has only just begun.

5. A Deeper Investigation into Condition Specific Core Promoter Elements

Elizabeth Rach conceived and performed all of the work in this chapter, except the position overrepresentation motif finder FREE was developed and published by Ken Yokoyama from Dr. Greg Wray's lab (Yokoyama, Ohler, and Wray 2009), and Stoyan Georgiev from Dr. Uwe Ohler's lab wrote the Pearson Correlation code that Elizabeth Rach used for the motif comparisons. The results are preliminary and have not been published.

5.1 Introduction

The spatiotemporal expression of genes is fundamental to proper functioning and development of multicellular organisms. The mis-regulation of genes at spatiotemporal conditions in the eukaryote *Drosophila melanogaster* can lead to malformation and death (Aoyagi and Wassarman 2001; Casares and Mann 1998). Transcription initiation is essential in the spatiotemporal regulation of genes. For eukaryotic protein coding genes and some regulatory RNAs, transcription is initiated through the ordered assembly of a pre-initiation complex (PIC) at the DNA (PIC) (Gross and Oelgeschlager 2006). The PIC contains the holoenzyme RNA polymerase II that is able to synthesize RNA and proofread transcripts. The second component of the PIC is a set of 5 general transcription factors (TFs): TFIIB, -D, -E, -F, and -H, that recognizes and unwinds the promoter DNA. The TFIID contains the well-studied TATA Binding

Protein (TBP). The third component of the PIC is a mediator that relays information from TFs bound to regulatory DNA sequence motifs to the polymerase (Boeger et al. 2005). TFs bind to regulatory sequence elements that can be found in trans up to several kilobases (kb) from the site of initiation and in the core promoter, the sequence approximately +/- 50 bp directly surrounding the transcription start site (TSS) (Latchman 2005). The repertoire of TFs in the nucleus during transcription initiation varies across developmental stages, cell types, and body parts (Schug et al. 2005). Upon the assembly of the RNA pol II at the TSS, the C terminal domain of the RNA pol II complex is phosphorylated, released from the core promoter, and transcription initiation is completed (Orphanides and Reinberg 2002).

To experimentally identify TSSs, 5′ capping data was used. With this capping technique, tag mappings were produced that spread over large regions of DNA located upstream of translational start codon(s) (Carninci et al. 2006; Kimura et al. 2006; Stapleton et al. 2002). Distinct TSSs were not clearly identified, and transcription factor binding sites (TFBS) were searched for using entire stretches of upstream sequence (Krivan and Wasserman 2001). As a result, associations of core promoters to condition specific profiles were often made on a whole gene basis (Parisi et al. 2004; Schug et al. 2005; Zhan et al. 2007).

In Chapter 3, we applied a hierarchical clustering technique to 5′capped Expressed Sequence Tags (ESTs), 200-500bp tags that map to the most 5′ of transcripts,

in *Drosophila melanogaster*. This method distinguished the most highly focused TSSs from dispersed transcription initiation and noisy 5′ background data (Juven-Gershon et al. 2008). Genes with alternative TSSs were discovered across the fruitfly genome on a large scale, and not only in isolation (Chen et al. 2007). In addition, Shannon entropy and tiling array data identified canonical TSSs having independent spatiotemporal profiles (Manak et al. 2006). An extensive genome wide mapping of the associations of TSSs to specific conditions was made that greatly increased our knowledge of the number of genes with core promoters driving specific expression patterns (Smith and Wakimoto 2007). The genetic components guiding these associations have been investigated in depth in a few instances (Corbin and Maniatis 1989; Margolis et al. 1995; Margolis et al. 1994). However, our understanding of the precise utilization of these nuclear proteins on a large scale is minimal.

The two major players of transcription initiation in the core promoter are the RNA pol II complex and the DNA sequence to which it is recruited. Numerous TATA-binding Associated Factor (TAF) proteins have been found across species that assist in the binding of the RNA pol II transcriptional machinery to the DNA at specific spatiotemporal stages (Green 2000; Hochheimer and Tjian 2003). In *Drosophila*, AT tracts are often bound in the minor groove of DNA by condition specific TAFs containing protein motifs called AT hooks (Aravind and Landsman 1998). AT hook motifs work in cooperation with other DNA binding proteins and can facilitate changes in the structure

of DNA. In recent work, Metcalf and Wassarman used gel mobility shift assays to show that TAF1 has an AT hook motif that binds to two AT tract sequence motifs, with a preference for the AAT sequence, in testes specific genes in *Drosophila* (Aravind and Landsman 1998; Metcalf and Wassarman 2007). The exact length and nucleotide composition of the TAF1 AT tract sequence motif remain to be found.

Wright and Wassarman also illustrated the role of TAF4 and TAF(II)250, respectively, in the positioning and stabilizing of the TFIID (Wassarman and Sauer 2001; Wright, Marr, and Tjian 2006). Hiller demonstrated that the gene *cannonball* encodes a dTAFII80 homolog that is important during male gametogenesis (Hiller et al. 2001), and that *nht, mia, sa,* and *rye* are four TAFs expressed in primary *Drosophila* spermatocytes (Hiller et al. 2004). In humans, Deato and Tjian showed that TAF3 was integrated into the TFIID complex during the differentiation of myoblasts into myotubes (Deato and Tjian 2007). In spite of these findings, experimental and computational evidence for the existence of the condition specific TAF binding sites in the DNA is very limited. The short sequence length and the large amount of nucleotide degeneracy of binding sites make them difficult to identify.

The pause button is a recently discovered core promoter motif that has been associated with the stalling of the RNA pol II during transcription initiation. The binding behaviors of RNA pol II to the DNA can be divided into 3 categories: active, stalled, or not occurring (Zeitlinger et al. 2007). When the RNA pol II is actively bound

to a gene, complete transcripts of the gene are produced. ChIP-chip assays identify active RNA pol II binding by observing uniform profiles of Pol II levels across the entire length of the transcript. When the RNA pol II is stalled, it is bound to the core promoter DNA near the TSS. However, it does not get released from the core promoter, and no transcripts of the gene are made. This behavior can be characterized on ChIP-chip assays by high concentrations of RNA pol II close to the TSS, with an absence of RNA pol II throughout the rest of the transcript. The third category of RNA pol II binding behavior is when it does not occur. The DNA is wound tightly in a double helix, and no transcripts of the gene are produced. When no RNA pol II is bound to the DNA, ChIP-chip assay profiles show an absence of RNA pol II across the whole genomic region.

The pause button motif was identified in motif searches of core promoters that exhibited stalled RNA pol II behavior in *Drosophila melanogaster* embryos, as revealed by ChIP-chip assays (Zeitlinger et al. 2007). The stalling behavior is believed to be embryo specific. However, as the assays were solely performed in embryo derived cell lines, additional analysis is needed across conditions to determine the scope of condition of the RNA pol II stalling. The consensus sequence of the motif was established as KCGRWCG. The combination of the INR, GAGA, DPE, and pause button motifs were predicted to contribute to the stalling of the RNA pol II. The pause button motif was found to occur in the window (-5,+55) from the TSS, with the most frequent positions occurring at (+20,+30) (Hendrix et al. 2008). A complete investigation of possible

sequence variants and location preferences of the pause button motif have not yet been explored.

One of the most common algorithms used to search for binding site motifs is expectation maximization (MacIsaac and Fraenkel 2006). In this model, the position weight matrix (pwm) and locations of each motif are estimated from missing data until the probability of a sequence converges to a maximum value. Significant deviations from the background model are deemed true motifs. Two vital motifs in *Drosophila melanogaster*, the TATA and INR, have been found using this algorithm (Ohler et al. 2002). Recent work has shown that incorporating additional genomic features, such as the clustering of motif locations and the conservation of sequence across species, into probability models improves the accuracy of motif prediction (Alkema et al. 2004; Blanchette et al. 2006; Pierstorff, Bergman, and Wiehe 2006; Siddharthan, Siggia, and van Nimwegen 2005).

The usage of position overrepresentation information has also proven to be a valuable technique in evaluating motif significance (Berendzen et al. 2006; FitzGerald et al. 2006). The experimental construction of a super core promoter has shown that the inclusion of multiple core promoter motifs at specific positions in *Drosophila* leads to an increased level of transcription (Juven-Gershon, Cheng, and Kadonaga 2006). An estimated 10 bp are required for 1 complete turn of the DNA helix (Trifonov and Sussman 1980). This implies that motifs spaced 10bp apart bind TFs on the same side of

the tertiary helical structure. Having this TF arrangement may increase transcriptional efficiency by increasing the rate of TF assembly, or the ability of the TF to recruit other components to the DNA. This arrangement may also allow physical interactions between the TFs and the DNA that further stabilize TF binding and increase the rate of transcription.

TFBSs can be classified into two groups based on their location in the DNA. The first group consists of motifs that have a defined spacing preference from the TSS. In *Drosophila melanogaster,* the most well-studied examples of this are the TATA, INR, DPE, and MTE (Juven-Gershon, Cheng, and Kadonaga 2006; Ohler et al. 2002). The second group of motifs has the ability to be located throughout the core promoter, for example the DRE. The complete motif composition and size of both groups is not known. Condition specific binding sites have not been identified on a large scale, and search techniques applied to core promoter data have used discrete, independent windows of pre-determined width for sequence analysis that limit the precision and scope of motif discovery (Berendzen et al. 2006; FitzGerald et al. 2006). Thus, the importance of TFBS spacing to the TSS remains unclear.

While advancements in our understanding of transcription initiation have been made, our knowledge of the genetic mechanisms governing the condition specific binding of TFs in the core promoter is in its infancy. In this work, we identified condition specific TFBS candidates using a statistically rigorous location

overrepresentation search algorithm. We compared our results to experimental and computational motif sources and analyzed core promoter sequence properties in a hierarchical fashion. Lastly, we classified motif variants of the pause button into a core promoter family.

5.2 Materials and Methods

5.2.1 Motif Searches and Clustering

An earlier version of the motif identification program FREE than that which was published was used in the motif searches. FREE was applied to 1,446 embryo and 673 head condition specific core promoter sequences extending 60bp upstream and 40bp downstream of TSSs identified in Chapters 3 and 4 (Rach et al. 2009). The (testes, ovary, larva/pupa, and Schneider cell) libraries each had less than 250 sequences, making the sets too small to produce accurate search results. Each position within the 100 nucleotide windows was considered a site, totaling 100 sites across all sequences. Statistically overrepresented oligos 6bp in length were identified that had at least 5 occurrences at a single site, and at least 30 occurrences across all sequences and all sites. Oligos were grouped into clusters using the Kullback-Leiber divergence cutoff of 0.4, and cluster motifs were determined using IUPAC nomenclature for the best consensus alignment, allowing for complete nucleotide degeneracy. The statistical significance and Gaussian parameters of the most significant oligo in the cluster was assigned to the

cluster motif. Clusters not having at least one 6mer with a p value less than $1e^{-10}$ were removed from the analysis.

Cluster motifs having Pearson Correlation (PC) < .8 to all other cluster motifs within each library were declared cluster group motifs. The statistical significance and Gaussian location information of the most significant cluster motif was inherited by the cluster group motif. Cluster motifs having PC $\geq$ .8 to other cluster motifs in each library were hand clustered. This was done by grouping all cluster motifs with PC $\geq$ .8 together and taking the longest motif as the representative. If all of the cluster motifs were the same size, the most significant one was chosen as the representative. Cluster motifs that were reverse complements and occurred at similar locations as the representative were discarded. Cluster motifs with +/- 1 the number of Gaussians located in similar positions as the representative and at most 1 interior or edge nucleotide mismatch from the representative, were incorporated into a cluster group motif consensus sequence. Cluster motifs that did not satisfy these criteria were not set aside. The process was repeated on these cluster motifs until all of the cluster motifs were grouped into cluster groups. A few assignments during hand clustering were done by eye.

Pairwise comparisons were made with cluster group motifs having PC $\geq$ .8 across embryo and head libraries, a difference of 1 Gaussian, and similar Gaussian locations. One cluster group motif from the library with the greater significance was chosen as the

representative of both libraries. References to FREE motifs in the text are cluster group motifs.

MEME searches were performed on 1,446 embryo, 673 head, and 222 testes specific sequences. Overlapping core promoter sequences for genes with more than 1 TSS were joined together in pairs. Genes with more than 2 overlapping core promoters accounted for less than 5% of all sequences, and did not greatly affect the results. MEME was applied to search for any number of motifs per sequence, and used a first order Markov Model as background. In addition to the embryo, head, and testes sequences, the background nucleotide frequencies incorporated 43 Schneider cell, 31 larva/pupa, and 10 ovary specific core promoter sequences. MEME searches returned the 25 most significant motifs, 5-15 nucleotides in length. The maximum number of input sequences was increased to 200,000 for the embryo sequences, and search results with e-values $\leq e^{-005}$ were deemed motifs, which resulted in the identification of 10 additional motifs. MEME motifs were not grouped into clusters or cluster groups.

The total set of 35 MEME motifs were compared to FREE results within each condition library. Those having PC $\geq$ .8 to a FREE motif were excluded from the set, reducing the number of motifs to 22. Motifs having PC $\geq$ .8 to more significant MEME motifs were also excluded from the set, reducing the number of motifs to 19. Search results were compared across libraries. In pairs having PC $\geq$ .8, the library in which the motif had a more significant p-value was chosen as the representative and the motif in

the library with the less significant p-value was excluded from the set. Sixteen MEME motifs satisfying these criteria were identified. We called motifs with a greater significance in one library than another 'condition specific motifs'.

5.2.2 Source Comparisons

Two methods were implemented to compare motif search results to known sources: consensus formatting and PC comparisons. In FREE searches, oligomer results were matched to IUPAC consensus sequences created from the position weight matrices for Ohler (Ohler et al. 2002), FitzGerald (FitzGerald et al. 2006), JASPAR (Bryne et al. 2008), Pause Button (Hendrix et al. 2008), and the 12 Genome Sequencing motifs (Stark et al. 2007). If a nucleotide had $\geq$ 50% frequency at a site, and the nucleotide was twice as common as the next most common nucleotide, the site was set to a single consensus. It two nucleotides made up at least 75% of the nucleotides at that site, double degenerate nucleotides were used. Triple or complete degeneracy was not considered.

In the STAMP comparisons, motif edged were trimmed if they had an information content < .4 (Mahony and Benos 2007). PC was used with an ungapped Smith-Waterman alignment to compare FREE oligomers to the FlyReg (Bergman, Carlson, and Celniker 2005), Fly (Bergman 2007), and Tiffin (Down et al. 2007) motifs. Motifs with e-value significances $\leq e^{-5}$ were considered matches. Cluster motifs cumulatively inherited all of the motif matches for each oligo in the cluster, and cluster

group motifs inherited all of the motif matches for each cluster motif. This included motif matches from less significant pairs having PC $\geq$ .8 across libraries.

As described in 5.2.1, MEME results were compared to FREE cluster groups using PC. Novel MEME motifs that did not match any FREE cluster group motifs were compared to the IUPAC consensus motifs using a PC cutoff of .8. The FlyReg, Fly, and Tiffin motifs were compared to the MEME results using STAMP as stated. MEME motif representatives inherited source matches for pairs across libraries having PC $\geq$ .8. Source matches to MEME motifs with PC $\geq$ .8 in the same library were excluded from the analysis.

5.3 Results

5.3.1 Distinct Core Promoter Motifs Are Identified Across Conditions

5.3.1.1 Motifs Identification by FREE and MEME

There were 123 motifs identified in *Drosophila melanogaster* from 2,425 condition specific core promoter sequences (see Table 6, Appendix F). In Chapter 3, 5′ capped Expressed Sequence Tags (ESTs), 200-500bp sequences that map to the most 5′ of transcripts, were hierarchically clustered to elicit 5,836 of the most highly utilized TSSs (Rach et al. 2009). Shannon entropy showed that 2,425 of the core promoters were utilized in specific body parts (head, ovary, testes) and time periods (embryo, larva/pupa, schneider cells) (Rach et al. 2009). As this is the most extensive TSS condition

specific profiling in *D.melanogaster*, we used this condition specific dataset as the foundation for the motif searches.

Table 6: Motif Identification Pipeline

Approximately 2,341 core promoters utilized in the embryo, head, and testes conditions were extracted in *Drosophila* melanogaster in Chapters 3 and 4 (Rach et al. 2009). FREE was implemented on these sequences according to the parameters in the Materials and Methods. It identified 368 oligomers that were 6 bp in length. Oligos having a Kullback Leiber divergence less than 0.4 were grouped into 196 clusters. Then, clusters having Pearson correlation coefficients greater than 0.8, reverse complement sequences, or different numbers of overlapping Gaussians were organized into 120 cluster groups by hand. FREE cluster groups were compared within and across conditions and those having Pearson correlation greater than 0.8 to a more significant cluster group within or across conditions were removed. This produced 107 unique cluster groups. MEME was applied to the 2,341 core promoter sequences utilized in the embryo, head, and testes, and MEME motifs that had a Pearson correlation coefficient greater than 0.8 to a FREE oligo were removed from the results. Overall, MEME found 16 additional motifs. Cumulatively, the FREE and MEME searches identified 123 motifs.

Number of:	Embryo	E & H	Head	Testes
Condition Specific Core Promoter Sequences	1,446		673	222
Oligos Identified Using FREE	269		99	0
Clusters	126		70	0
Cluster Groups	77		43	0
Unique Cluster Groups Across Conditions	63	11	33	0
Additional Motifs Identified Using MEME	8	3	4	1
Total Number of Motifs Identified = 123	71	14	37	1

We applied two different search techniques to the data (see Materials and Methods). The first program called FREE advances current methodologies by implementing smooth functions across continuous window spaces. It is based on a motif location function (MLF) that measures the positional overrepresentation of an oligomer with respect to a reference point. The MLF is a composite of a baseline function and a signal function. The baseline function measures a motif's nucleotide frequency as a background model. The signal function, H(x), measures motif position deviations from the background. The signal function is a linear combination of un-normalized Gaussian motif peak locations. A log likelihood ratio test is used to determine the parameters of the Gaussian peaks, and a F-test is used to measure the statistical significance of the MLF that best fits the data. Rigorous linear regression statistics are used to return significant oligomer sequences and their corresponding MLF parameters. Then, FREE uses Kullback-Leiber divergence to cluster oligomers, of user specified length n, with similar sequence and position (Yokoyama, Ohler, and Wray 2009).

The second algorithm that we used to identify condition specific core promoter motifs was expectation maximization in the search program MEME (MacIsaac and Fraenkel 2006). This well established search technique does not require motifs to have a positional overrepresentation. Thus, the results of MEME complement FREE and provide a more complete investigation of condition specific motifs in the core promoters.

The sequence searches returned 123 core promoter motifs (see Table 6, Appendix F). MEME found 29 unique motifs, 13 of which had matches to FREE sequences. This reduced the overall number of additional MEME motifs to 16. FREE identified 107 motifs, 87% of the total number discovered. There were 94 of the FREE motifs found to be mutually exclusive from the MEME motifs, resulting in nearly 6 times more motifs than MEME. This ratio reflects the difference in search techniques and emphasizes the importance of proper spacing of core promoter motifs for efficient utilization by the RNA pol II and its associated factors.

5.3.1.2 Condition Associations

The 123 discovered motifs were distributed across the embryo, head, and testes specific libraries (see Table 6, Appendix F). The number of motifs returned in the embryo and head libraries scaled according to the number of sequences used for input. Only 14 of the motifs were found in both the embryo and head libraries. This accounts for approximately 11% of the total number of motifs. As dual detection required similarity in nucleotide composition, as well as, location (see Materials and Methods), the low percentage of shared motifs indicates the existence of binding site sequence differences between the condition specific core promoters.

Figure 43: Logo of the testes specific core promoter motif consensus sequence

MEME was run on 222 testes specific core promoter sequences. The position weight matrix of Testes Motif 1 was converted to a consensus sequence using the IUPAC code and double degeneracy, as for the source comparisons (see Materials and Methods). Then, the consensus sequence was input into STAMP, and the logo was created. Edge positions having less than 0.4 bits were removed from the logo (Mahony and Benos 2007).

One motif was found in the testes specific core promoters using MEME (see Materials and Methods, Figure 43). The motif was selectively not found in the embryo or head FREE and MEME searches. The testes motif is 12 nucleotides in length and consists of an AAAT sequence at nucleotide positions 1-4 and 7-10. The nucleotides T and C share a common frequency at position 6. With the occurrence of a T at position 6, positions 1-9 are palindromic, making the motif non-directional.

5.3.1.3 Existence of Motif Location Hotspots

For all conditions, FREE returned location information regarding the amplitude, mean, and standard deviation of each Gaussian used to model motif positions. The number of Gaussians per motif ranged from 1 to 5, with 29 motifs having 1 Gaussian, 36 motifs having 2 Gaussians, 29 motifs having 3 Gaussians, 9 motifs having 4 Gaussians,

and 4 motifs having 5 Gaussians. The most common number of Gaussians per motif was 2.

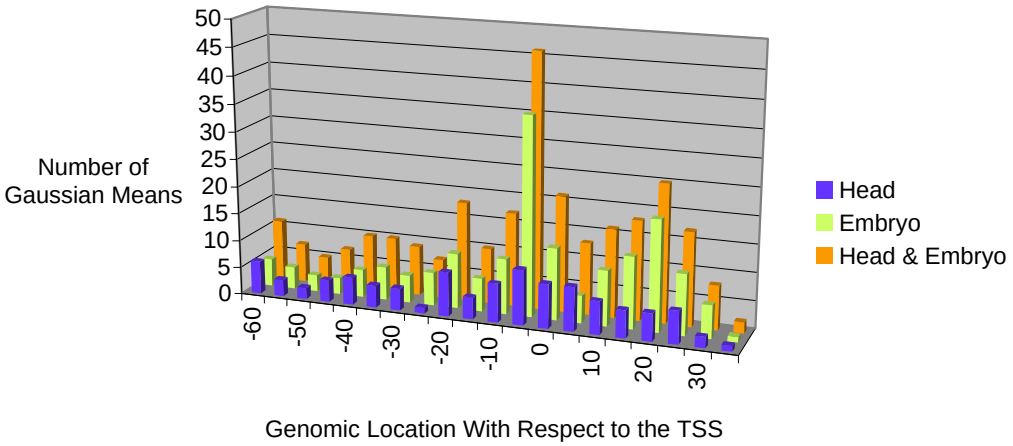


Figure 44: Motif Locations Reveal Condition Specific Hotspots

The 100bp sequence of the core promoter was divided into 5bp windows. The position overrepresentation information for each FREE cluster group was chosen from the most significant cluster. The number of Gaussians modeling the position overrepresentation of each cluster group ranged from 1 to 5. The mean, μ, of each Gaussian was binned into one of the 5bp windows according to condition under which the cluster group is utilized. The frequency of Gaussian means was summed across conditions to identify shared locations containing motif variants.

Figure 44 shows that the frequencies of Gaussian means, μ, used to model motif locations are spread unevenly across conditions and throughout the core promoters. The number of Gaussian means for embryo specific core promoters shows a significant enrichment in 2 peaks, (-5,0) and (+20,+25). Head specific promoters reveal a larger, less

striking region of utilization from (-5,+25). Cumulatively, a third peak of Gaussian means occurs in the embryo and head set at (-20,-15) that was not present in either of the core promoter sets individually. Previous work identified motifs in all three hotspot regions (FitzGerald et al. 2006; Ohler et al. 2002). The regions may be functionally important because of the unique motifs that they contained. Figure 44 shows that a multitude of additional motifs exist across core promoters at these locations. This suggests a transcriptional role of the locations beyond that of individual motif contributions.

5.3.2 Comparisons to Alternative Motif Sources

5.3.2.1 Comparative Overlap

Computational and experimental sources were used in comparisons with the 123 condition specific core promoter motifs. The set of computational resources included 10 motifs from previous Ohler analysis (Ohler et al. 2002), 4 motifs from FitzGerald's work (FitzGerald et al. 2006), 35 motifs from the 12 genome sequencing project (Stark et al. 2007), 120 motifs from the TIFFIN database (Down et al. 2007), and 1 motif from pause button analysis (Hendrix et al. 2008) (see Materials and Methods). The experimental sources consisted of 13 motifs from the JASPAR (Bryne et al. 2008), 87 motifs from FlyReg (Bergman, Carlson, and Celniker 2005), and 62 motifs from Fly (Bergman 2007). All together, this totals 332 motifs used in comparison with each of the 123 identified

motifs: 170 computationally identified motifs and 162 experimentally derived motifs. However, the existence of motif repertoire overlap across sources and methods, reduces the total number of distinct motif comparisons.

More than 87%, or 107 out of the 123 motifs, had at least one match to a motif in one of the comparative resources. Approximately 50%, or 62 out of the 123 motifs, had at least one match to motifs in both computational and experimental sources. The Gaussian means of the most significant motif matches correlate to known locations for 9 out of 12 of the Ohler and FitzGerald motifs.

5.3.2.2 Normalized Comparisons

The overlap of identified motifs to matches in comparative sources varies greatly. Two measures of coverage across sources were calculated, as seen in Figure 45. The first metric measures the percentage of identified motifs that have matches to source motifs. The Tiffin and 12 genomes sequencing motifs had the greatest percentages of overlap, followed closely by FlyReg. The Pause Button and FitzGerald motifs had the lowest percentages of motif support.

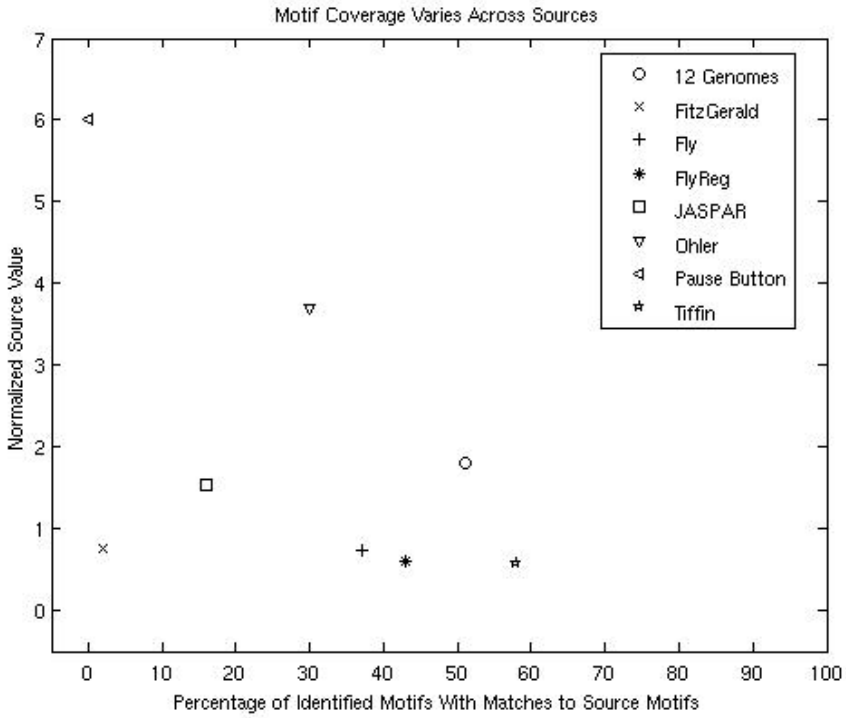


Figure 45: Comparison of Motifs to Sources

A plot comparing the overlap of identified motifs to matches in alternative sources. Motif matches were found through the use of consensus sequences in FREE, Pearson correlation comparisons, and STAMP (see the Materials and Methods). Motif matches were cumulated across the embryo, head, and testes libraries. The percentage of identified motifs with matches in each source was calculated by dividing the total number of motifs identified that had matches to motifs in the source, by the total number of motifs identified, 123. A normalized source value was obtained by dividing the total number of motifs identified with matches to motifs in the source by the total number of motifs in the source.

With the percentage measurement, the amount of motif overlap scales according to the number of motifs in the source. To account for this bias, a second measure was created that normalized the number of motif matches by the size of the source (see Figure 45). As more than one identified motif could match one motif in the source,

normalized values greater than 1 could result. When normalized, TIFFIN received a low value on account of the large size of the database and the high number of motifs that did not have matches to the identified motifs. The pause button, Ohler, and, 12 genomes sequencing sets had the highest measure of similarity to the identified motifs. This can be expected from the pause button motif as addressed above, and from the Ohler dataset, as it was the foundation of the core promoter localizations in Chapters 3 and 4 (Rach et al. 2009).

The 12 genomes sequencing set received high scores with both numerical comparisons to the identified motifs, in spite of the reduced number of species used for identification. The sequencing motifs were produced from conservation analysis across 12 species, while the FREE motifs were found from statistical analysis of precise overrepresentation of 6-mers in 1 species.

5.3.2.3 Presence of Novel Motifs

Overall, 16 of the 123 motifs did not have matches to any of the motifs in the comparative sources (see Table 7). Of the 16 motifs, 68% were identified in the embryo, and 32% were found in the head. All but 1 of the motifs was found using FREE. Consensus sequences showed ample nucleotide variation and motif lengths were of expected size. Head_MOTIF8 was longer than the most common FREE motifs but, within range of the MEME results. The number of Gaussians for each motif ranged from

1 to 3, and the Gaussian means were spread across the 100nt core promoter window. These features are characteristic of the identified FREE motifs.

Table 7: Novel Motifs

Of the 123 motifs identified in the FREE and MEME searches, 16 motifs did not have any matches to motifs in the Ohler, FitzGerald, 12 genomes, Tiffin, Pause Button, JASPAR, FlyReg, or Fly sources. The first, second, and third columns designate the search method, consensus sequence, and library in which the motif was identified, respectively. The fourth column lists the p-value of the most significant MLF in the FREE searches. The last column gives the parameters of each Gaussian in the MLF. As MEME searches do not return location information, this information is Not Available (N/A) for the Head-MOTIF8.

Search	Consensus Sequence	Library	P-val	Gaussian Parameters (amp, mean, sd)
FREE:	ATCATT	Embryo	1.30E-51	(445.23,-3.36,0.5)
				(316.62,0.28,0.53)
FREE:	CGTCAG	Embryo	1.90E-46	(550.28,-4.44,0.53)
				(135.12,-48.00,<.5)
FREE:	GATTCA	Embryo	3.00E-43	(493.62,-5.67,<.5)
				(136.82,-12.00,<.5)
FREE:	CCCTGG	Embryo	3.20E-28	(26.82,-25.00,<.5)
				(19.83,14.00,<.5)
				(12.98,-22.00,<.5)
FREE:	CGGAGC	Embryo	4.60E-28	(270.14,26.00,<.5)
				(131.83,15.00,<.5)
FREE:	GACAGT	Embryo	1.10E-33	(425.77,-3.92,<.5)
FREE:	AGTCGA	Embryo	1.50E-27	(535.27,-0.28,0.60)
				(132.86,-43.00,0.77)
FREE:	AGCCAG	Embryo	7.70E-27	(268.70,-4.00,<.5)
				(130.39,-57.00,<.5)
				(130.39,10.00,<.5)
FREE:	CAGTAT	Embryo	8.70E-57	(632.11,-1.48,<.5)
				(136.07,4.00,<.5)
FREE:	AGCAAC	Embryo	6.10E-30	(313.48,-23.28,0.54)
				(255.02,25.64,0.74)

				(187.81,7.51,0.55)
				(133.16,16.00,<.5)
FREE:	AAGCAG	Head	1.90E-42	(83.88,-27.55,<.5)
				(30.18,-17.00,<.5)
FREE:	TCGCTT	Head	9.90E-18	(57.73,2.00,<.5)
FREE:	CTGGTC	Head	1.20E-21	(57.73,-6.00,<.5)
				(28.02,-30.00,<.5)
FREE:	GTTGAA	Embryo	2.70E-13	(392.13,0.97,<.5)
FREE:	GTTGAC	Head	1.20E-52	(68.27,29.46,0.63)
MEME:	TCGGCMAGCCATCGT	Head_MOTIF 8	N/A	N/A

In previous applications, FREE exhibited 70% accuracy when verifying known JASPER motifs at p-value threshold $1e^{-10}$ (Yokoyama, Ohler, and Wray 2009). With this level of accuracy, 32 of the 6-mers identified in these FREE searches are estimated to be false positives. The 15 FREE motifs without matches may account for 47% of the total number of predicted false positives.

5.3.2.4 Existence of a Pause Button Family

A family of 6 motifs matched the pause button consensus sequence found in previous work (Hendrix et al. 2008). Five out of six of the motifs were found selectively in the embryo, and the sixth motif, TGCGWTCRGTT, was identified in both the head and embryo core promoter sequences. This agrees with the selective embryonic utilization of the pause button (Hendrix et al. 2008). A logo of the consensus sequence of the 6 motifs was created in STAMP (see Figure 46A). The logo has the same CG di-nucleotide sequence repeat and spacing as the original pause button motif found in

previous work, but differs in the first two nucleotides. Positions 4 and 7 are dominant in the motif alignment, and are surrounded by nucleotides having second order degeneracy with thymine, T.

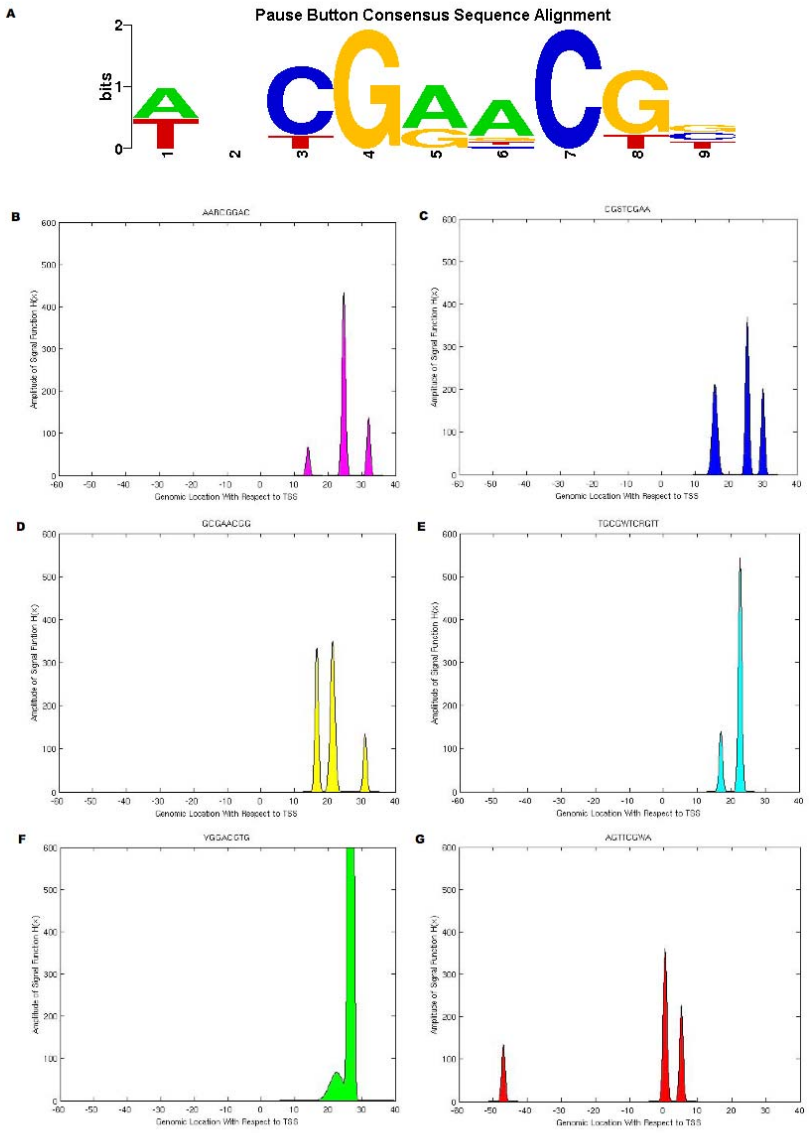


Figure 46: Pause Button Matches

Six of the cluster groups identified in the FREE searches had matches to the pause button motif: AGTTCGWA, CGSTCGAA, GCGAACGG, TGCGWTCRGTT, YGGACGTG, AARCGGAC. Cluster group matches to the pause button motif were inherited from all clusters (see Materials and Methods). A logo of the consensus sequence of the 6 cluster group motifs was made in STAMP (Mahony and Benos 2007). Each of the 6 motifs were weighted equally when creating the logo, and edges of the logo having information content less than 0.4 were trimmed. B-G Graphs of the signal function H(x) of the 6 cluster groups matching the pause button motif. The parameters of each signal function H(x) are used to plot the position overrepresentation of the cluster group according to its genomic location with respect to the TSS, and the amplitude of the Gaussians. The parameters of each signal function for a cluster group are inherited from the most significant cluster (see Materials and Methods). The area under each signal function is shaded accordingly. No maximum constraint was placed on the height of the amplitudes. The height of the second Gaussian centered at 26.66 used to model the overrepresentation of YGGACGTG is 7958.30 (see Figure 46F). It is the largest peak of all the Gaussians plotted, and is cut off at 600 to maintain the visual uniformity of the graph parameters of the 5 other signal functions.

The positions of 5 out of 6 of the motifs in the pause button family were modeled with signal functions within 10bp in proximity, in the window (+15,+35), downstream of the TSS (see Figure 46B-G). The location of this window overlaps that of the DPE, and the size of the window doubles initial estimates (Hendrix et al. 2008). The signal function of AGTTCGWA showed the greatest divergence from the common window, with all of its Gaussian peaks occurring upstream of +10bp from the TSS. Half of the signal functions are modeled with 3 Gaussians, each having a dominant center peak reaching approximately 400.

5.4 Discussion

In this work, we defined condition specificity as the selective expression of a characteristic pattern of a gene during a single spatiotemporal state, namely the embryo, head, and testes. This separation of conditions was precise enough to detect changes in the initiation of transcription at different locations. Condition specific motifs in the core promoter can be divided into 2 groups: motifs utilized based on position overrepresentation, and those utilized, regardless of location from the TSS. Here, FREE returned 107 motifs having a position overrepresentation from the TSS. This large number highlights the importance of proper spacing for the transcription initiation machinery. It may result from the dependence of the sequence motifs on the organization of the proteins within the RNA pol II complex. The RNA pol II contains hundred of proteins and TFs that are systematically arranged. When TFs recruit the RNA pol II to the DNA, precise alignments of core promoter motifs efficiently bind to corresponding protein motifs in the RNA pol II. As a result, the distances between the core promoter binding sites may mirror the distances between protein motifs in the RNA pol II complex.

The spacing of core promoter motifs may also arise from changes in conformation induced by TFs. The binding of TFs to the DNA can be considered a dynamic process because TFs can cause conformational changes in the DNA that alter its structure (Westwood and Wu 1993). DNA previously inaccessible to TFs can be made

available and DNA once available can become unavailable. The existence of core promoter motifs at these locations can stimulate TF binding, and augment the rate of transcription initiation. When the TBP binds to the TATA box, conformational changes are induced. This causes additional TFs to be recruited to the DNA that enhance the rate of transcription initiation (Latchman 2005). Thus, precise core promoter motif spacing can arise in coordination with conformational changes.

While numerous position overrepresented core promoter motifs were identified in the embryo and head, there were few motifs found present in both conditions. This can be caused by promoters having divergent sequence composition across conditions, or by motifs with similar sequence composition having different overrepresented locations in the FREE searches. The latter explanation only accounted for a few cases, implying that the large number of condition specific motifs was found from a difference in motif sequence composition. Work by colleagues previously noted that core promoters utilized under more specific conditions were associated to lower CpG nucleotide frequencies (Schug et al. 2005). Here, we show that sequence differences exist on a higher order than with individual di-nucleotide frequencies. Entire motif repertoires differ across the embryo, head, and testes libraries. These condition specific sequences may enable the differential regulation of transcription when the same TFs are present across conditions. This occurs with with Pit-1. When Pit-1 is bound to a Prolactin promoter, transcription is induced. However, when Pit-1 binds to a different

sequence motif in the promoter, the co-repressor NCo-R is allowed to bind, and transcriptional repression results (Marx 2000).

The existence of distinct core promoter motifs in the embryo, head, and testes libraries may also accommodate the binding of specific TAFs not present under all conditions (Wassarman and Sauer 2001). Chapter 4 showed that alternative core promoters for one gene can be differentially utilized under spatiotemporal conditions (Rach et al. 2009). The differences in condition utilization and usage of alternative core promoters create a competitive landscape within the domain of one gene's upstream sequence. Under these pressures, core promoters must maintain motifs commonly utilized across conditions, while at the same time preserve those that are selectively utilized under rare, or specialized conditions. This creates a delicate balance that must be maintained between general transcription and condition specific core promoter utilization.

The testes specific motif identified in this work satisfies the nucleotide composition and length of an AT track sequence. The motif may work in cooperation with TAF1, or another testes specific TAF, to facilitate the faster assembly of the RNA pol II, or increase its stability. The position overrepresentation information showed stark differences across the embryo and head conditions. Embryo specific motifs revealed 2 peaked hotspots, while the head specific motifs were more smoothly spread across a larger window. This difference in motif locations may result from the occurrence of

epigenetic changes that limit core promoter utilization. A peak of Gaussian means occurs at locations (-20,-15) cumulatively in the head and embryo conditions, but not in either condition individually. This hotspot region may serve as a binding site for TFs expressed generally across both conditions, rather than selectively in one condition. Alternatively, it may be a position specific hotspot in the embryo that requires a larger number of tags in order to be more clearly distinguished.

In addition, we observed a low overall frequency of motif variants at the TATA box (-32) location and at locations upstream of the TSS. Only 25-30% of core promoter sequences are known to have TATA boxes (Kutach and Kadonaga 2000; Ohler et al. 2002). This has led scientists to search for alternative motifs to the TATA. The lack of motif variants at the (-32) location may signal an avoidance of motif interference with the important functioning of the TATA. It could also result from a greater usage of core promoter motifs downstream of the TSS. The downstream motif variants may help to define the direction of transcription by properly orienting the RNA pol II during assembly, similar to that of the BRE and TATA motifs (Littlefield, Korkhin, and Sigler 1999). The core promoter architecture may not directly affect the assembly of the RNA pol II, but make the DNA more accessible downstream of the TSS, providing a natural orientation for the direction of transcription. A large number of the identified core promoter motifs showed matches to enhancer sequences in the Fly, FlyReg, and JASPAR databases. Enhancers are known to have the same sequence composition as core

promoter motifs, although, they typically act at a distance from the TSS (Latchman 2005).

The pause button is a recently discovered motif that is associated with the stalling of the RNA pol II during embryonic transcription. Here, we identified 6 motifs with matches to the pause button consensus sequence. This extends the number of pause button motifs from 1 consensus sequence to a family of 6 possible sequence variants. Only 1 out of 6 of the motifs were identified in the head library. This result supports the embryo specific utilization of the pause button. Five out of 6 motif variants were located in the (+15,+35) window from the TSS. The motif located outside of the preferred window may be utilized under distinct or rare conditions, be a false positive, or have incorrect position information. Additional motifs and combinations thereof may play a role in the stalling of the RNA pol II that are not identified or discussed.

The exact biological function of the stalling is not known. It has been suggested that RNA pol II stalling actively represses the core promoters of genes utilized during later stages of development, similar to that of the GAGA (Hendrix et al. 2008; Lee et al. 2008; Zeitlinger et al. 2007). Alternatively, permanaganate assays in *Drosophila* embryos have suggested that the pause button functions to prepare genes for activation by responding to extracellular signaling molecules present during periods of rapid spatiotemporal change (Zeitlinger et al. 2007). Under both models, we presume that more than one pause button motif is required to differentially regulate the genes under

more than one stage of development or condition. The 6 pause button motifs identified here may uniquely correspond to the morphological stages of development, and their accompanying nuclear spatiotemporal environments. Utilization of these 6 motifs in combinations may further amplify the number of conditions under the control of the pause button.

While some condition specific differences in transcriptional programs have been known to exist, the initiation of transcription has long been considered a uniform and often sloppy process. Through the incorporation of condition utilization and position overrepresentation information, we have identified condition specific core promoter binding sites on a high throughput scale. We have shown the existence of systematic differences in sequence features across transcriptional programs, and suggested that condition specific transcriptional programs may have a greater impact on an organism's ability to properly develop and adapt than previously thought. As the identification of TSSs and their condition specific utilization increases with advancing technology, we expect to gain a deeper insight into the characteristics of the sequence features guiding the spatiotemporal regulatory code.

6. A Paired-End Sequencing Strategy to Map the Complex Landscape of Transcription Initiation

Elizabeth Rach prepared the *Drosophila* embryos and RNA that was used to provide the incentive for this study. She also performed the core promoter motif analysis, the chip-ChIP binding comparisons, and wrote the corresponding sections in this chapter. The experiments were performed by Dr. Ting Ni from Dr. Jun Zhu's lab and the identification of TSSs from clusters was done by Dr. David Corcoran from Dr. Uwe Ohler's lab. The work was in revisions with *Nature Methods* in March 2010.

6.1 Introduction

The transcription of a gene by RNA Polymerase II (Pol II) is a fundamental step in the regulation of gene expression in eukaryotes. To initiate and modulate transcription, regulatory factors interact with a variety of chromatin and DNA sequence features in regulatory regions. Central to the process of transcription initiation is the core promoter region of approximately 100nt centered on the transcription start site (TSS) of a gene. Within this region, sequence specific factors of the basal transcription machinery interact directly with the DNA to ensure the proper recruitment of Pol II. These transcription factors (TFs) recognize and bind to short, degenerate sequence motifs. Contrary to the simple picture in many textbooks, which often present the basal machinery as invariable, and that core promoters generally share the same motifs, many

recent studies have demonstrated the diversity in both basal transcription factor complexes as well as the features of genomic regions in which they bind (Juven-Gershon and Kadonaga 2009; Ohler and Wassarman 2010; Smale and Kadonaga 2003). We are still beginning to truly understand the diversity at the transcription initiation level, and how it provides for additional regulatory control of gene expression (Butler and Kadonaga 2001; Hochheimer et al. 2002; Holmes and Tjian 2000; Isogai et al. 2007; Juven-Gershon et al. 2008; Lee et al. 2005).

Methods to sequence 5′ complete transcripts at high throughput have provided the breakthrough for genome-wide identification of TSSs (Carninci et al. 2005; Shiraki et al. 2003; Suzuki and Sugano 2003; Tsuchihara et al. 2009; Zhang and Dietrich Mapping of transcription start sites in saccharomyces cerevisiae using 5' SAGE 2005). In particular, the capped analysis of gene expression (CAGE) protocol has been used to generate comprehensive mammalian libraries of short sequence tags. Aligning these tags back to the genome has led to high-resolution TSS maps, and helped to identify distinct transcription initiation patterns (Ahsan et al. 2009; Carninci et al. 2005; Carninci et al. 2006; Kodzius et al. 2006; Suzuki et al. 2009). In some promoters, transcription initiates from the same exact location; in others, it initiates more uniformly across wider genomic windows. Different sequence features have been found to be associated with these different patterns, such as an overrepresentation of the canonical TATA box sequence motif in ‘specific location’ promoters, and CpG islands overlapping ‘broad range’

promoters (Carninci et al. 2006). The majority of studies making use of the CAGE technology have been focused on mouse and human. There has not been an attempt thus far to investigate on a similar scale whether different initiation patterns can also be found in other animals, such as the fruit fly *Drosophila melanogaster*, and whether these are associated with distinct sequence features as well.

The CAGE protocol has recently been integrated with high-throughput sequencing technology. In particular, deepCAGE is based on the Roche/454 sequencing platform (Valen et al. 2009). While deepCAGE produces a large number of reads per sample, the total number of reads is on a smaller scale than what may be achieved by other platforms such as Illumina or SOLiD. In addition, deepCAGE produces a single, typically 20-nt-long sequence tag from the most 5′ end of the transcript, which may be too short to guarantee a unique and correct alignment to the genome, especially in the presence of sequencing errors. Such challenges could in theory be solved by longer reads or paired-end reads. The latter strategy is expected to be more advantageous because it can provide additional information on the local transcript structure.

We present a Paired-End Analysis of TSSs (PEAT) strategy, by which an individual TSS tag (20nt sequence from the most 5′ end of the transcript) is paired with a ~20nt downstream tag from the same gene. The PEAT method was applied to analyze capped transcripts of *D. melanogaster* mixed-stage embryos. We obtained 15 million mappable read pairs, collectively defining TSSs for more than 5,500 genes. Our results

further uncovered that *Drosophila*, like mammals, has multiple initiation patterns, each of which is associated with a distinct set of sequence motifs. Furthermore, we found that ~25% of 5′ capped reads align to the coding region of the *Drosophila* genome. Extending the previous findings in mammals (Carninci et al. 2006; Fejes-Toth et al. 2009), we provide strong evidence that these transcripts result from posttranscriptional modification rather than de novo transcription from the coding region. Together, these results demonstrate that PEAT is an improved strategy to map and characterize the landscape of transcription initiation in higher eukaryotes.

6.2 Materials and Methods

6.2.1 Paired-End Library Preparation

Mixed stage fly embryos (0-24h) were collected according to a standard protocol (Manak et al. 2006). We used TRIzol reagent (Invitrogen) to extract total RNA. RNeasy Mini kit (QIAGEN) was used for cleanup and on-column DNase I digestion to remove genomic DNA according to the manufacturer's protocol. 150 μg purified RNA was enriched for poly(A)+ RNA with Dynabeads Oligo(dT)25 (Invitrogen) according to a modified protocol (see Appendix G). 2 μg of poly(A)+ RNA was BAP/TAP (Bacterial Alkaline Phosphatase/Tobacco Acid Pyrophosphatase) treated and a chimeric linker tagged with a MmeI site was ligated to its 5′ end. The RNeasy MinElute kit (QIAGEN) was used to remove excessive chimeric linkers. Random primers tagged with a MmeI

recognition site were used to initiate reverse transcription. First strand cDNAs were amplified using 5 cycles of PCR, and the products were purified with DNA clean & concentrator-5 kit (ZYMO). Circularization was performed with a "collector" oligonucleotide, which converts the PCR product into a single-stranded circular DNA. After Exo I (NEB) and Exo III (NEB) digestion to remove linear DNAs, rolling circle amplification (RCA) was performed to amplify the remaining circular DNAs. The RCA products were digested with MmeI (NEB) to generate a specific 93~95bp band. The desired product was ligated with two Illumina Paired-End adaptors and amplified with low-cycle PCR. After size-selection and validation by Sanger sequencing, the final library was sequenced using an Illumina GAII with a paired-end module (see Appendix G).

6.2.2 Paired-End Sequencing and Read Mapping

Two technical replicates of the embryo library were sequenced as 36mers from each side using Illumina GA II. Before mapping, we filtered low-quality reads and short tags with unidentified linker sequences. The Novoalign short read aligner (v1.05.02; www.novocraft.com; *parameters:* score difference=30, report strategy='All') was used to align the paired reads independently to the *D. melanogaster* genome (FlyBase v5.14(Tweedie et al. 2009)). All alignments with up to one mismatch beyond their optimally aligned location for each read were collected. Since 3' reads might overlap a

splice junction, we mapped all those 3′ reads to the transcriptome where the 5′ read of a pair aligned uniquely, and the 3′ read did not map at all to the genome. Similar to the genomic alignment, we collected all 3′ read locations with one mismatch beyond the optimally aligned location. 5′ and 3′ read pairs that mapped in the same orientation within 200,000 nt on the same chromosome were flagged as 'aligned'. The cumulative Novoalign alignment score for both reads in the pair was used to classify the alignment specificity. Read transcript locations were classified into 6 possible categories based upon FlyBase: annotated TSS, ≤250nt upstream of an annotated TSS, 5′ UTR, coding region, 3′UTR, intron, and intergenic region. If a read could be classified into multiple categories because of overlapping transcripts, the read was assigned to one location based on the following priorities: (1) FlyBase annotated TSS, (2) 5′ UTR, (3) ≤250nt upstream of an annotated TSS, (4) coding region and (5) intron.

6.2.3 Transcription Start Site Cluster Identification

The feature density estimator F-Seq (Boyle et al. F-seq: A feature density estimator for high-throughput sequence tags 2008) (*parameters:* feature length=30, fragment size=0) was applied to the 5′ reads of the uniquely aligned pairs from both replicates in order to create a smoothened estimate of read distributions. A genome-wide background density estimate was calculated by taking the mean of F-Seq values sampled from across the genome, with each chromosome being sampled in proportion

to the number of reads aligned to it. Putative read clusters were defined as regions where the F-Seq value was greater than the background estimate. To eliminate lengthy tails in the distributions and create a robust cluster, we re-sized the clusters to the shortest distance that contained 95% of the reads. Clusters with tag numbers exceeding a stringent threshold (typically greater than 100 reads) were then considered as TSS clusters. Clusters were classified into different initiation patterns by the following definitions: NP clusters contained ≥50% of the reads within ±2nt of the mode and span <25nt; BP clusters were those that contained ≥50% of the reads within ±2nt of the mode and are ≥25nt in length; all other clusters were classified as WP. TSS cluster locations were determined according to FlyBase, similar to individual reads. If a cluster overlapped an annotated TSS, it was classified as such; otherwise, the classification was based on the mode of the cluster. If the mode fell into multiple categories because of overlapping transcripts, the cluster was classified according to the priorities listed in the previous section. To summarize the terminology, a TSS refers to a genomic location to which at least one 5′ capped sequence tag was aligned; a TSS cluster is a distinct region of TSSs above background; and the initiation pattern describes the distribution of TSSs within a cluster. In all cases, the mode of the cluster is used as the reference TSS for a cluster.

6.2.4 Core Promoter Motif Analysis

We considered the subset of TSS clusters overlapping an annotated TSS, ≤250nt upstream of a TSS, or in the 5′ UTR of a gene. The promoter sequences ±100nt surrounding the mode were extracted. The position weight matrix scanning program PATSER (Hertz and Stormo 1999) was applied to the plus strand of each sequence using pattern-specific background Markov models (Table 8). The relative frequency matrices of six previously described core promoter motifs (Ohler1, DRE, TATA, INR, Ohler6, Ohler7) (Ohler et al. 2002) as well as shortened non-overlapping matrices for the two motifs DPE and MTE used in Chapters 3 and 4 (Rach et al. 2009) were evaluated. All locations with a p-value ≤ 10^{-3} were deemed motif matches. Motif match counts were then binned into 5nt windows for each initiation pattern. To assess the background level of motif matches, the analysis was repeated on three sets of 1,000 random intergenic sites. The mean value within each bin was calculated. We define the preferred location for a motif as any 5nt window with a mean normalized count equal to or greater than 5-fold enrichment over background.

Table 8: Read Cluster Specific Background Markov Models Used in Identifying Core Promoter Motifs

	A	C	G	T
Narrow with Peak	0.262	0.242	0.236	0.260
Broad with Peak	0.280	0.222	0.216	0.281
Weak Peak	0.298	0.214	0.204	0.285
Coding Region Read Clusters	0.260	0.261	0.256	0.223
Random Intergenic Sites Set #1	0.292	0.203	0.207	0.297
Random Intergenic Sites Set #2	0.295	0.205	0.205	0.295
Random Intergenic Sites Set #3	0.299	0.202	0.202	0.297

6.2.5 ChIP-chip Transcription Factor Binding

We collected 612 TBP, 1,073 TRF2, and 298 TBP/TRF2 binding sites from Isogai *et al*.(Isogai et al. 2007), and converted the release 4 coordinates to release 5 using the FlyBase map coordinate converter (Wilson, Goodman, and Strelets 2008). For comparison, we selected read clusters further than 500nt from any other cluster; this was necessary because of the limited resolution of the ChIP-chip data. A read cluster was counted as being bound by one or both of the factors if the mode of the cluster was within 50nt of a binding site. Overall, we had 63 clusters bound by TBP, 432 clusters bound by TRF2, and 47 additional clusters bound by both TBP and TRF2. To account for

the different coverage of initiation patterns by ChIP-chip, the percentage of TF binding was calculated from counts normalized to the number of occurrences per 1,000 TSSs per 1,000 ChIP-chip binding sites and then divided by the normalized number of promoters with TF binding.

6.2.6 Identification of Novel Transcription Start Sites

We defined a novel transcription start site for a gene as a TSS cluster more than 250nt upstream of the most distally annotated start site according to FlyBase. Candidates for experimental validation were then required to contain a cluster with at least 100 reads, and with at least 80% of its 3′ paired reads mapping to a transcribed region of that gene. Novel 5′ exons identified in a previous analysis of whole-genome tiling expression arrays (Manak et al. 2006) were transferred from Release 4 to Release 5 of the *D. melanogaster* genome. We excluded from the analysis any exons that overlap a transcribed region as defined by FlyBase. Due to the more limited resolution of the tiling arrays, a TSS cluster was considered as overlapping one of the tiling 5′ exons if it fell within 50nt of that exon.

6.2.7 Experimental Validation of Novel TSSs and Internal Capped Transcripts

Two independent approaches, oligo-capping and cap-trapping, were used. For the cap-trapping method, total RNA isolated from 0-24hr fly embryo (0-24h) was

reversely transcribed with random hexamers and Superscript II reverse transcriptase. The resulting RNA/cDNA hybrids were oxidized with 10 mM NaIO4 in 66 mM NaOAc (pH 4.5) by incubation on ice for 45 min. Biotinylation was then carried out by adding 10 mM biocytin hydrazide (Sigma) and 50 mM sodium citrate (pH 6.1), followed by incubation overnight at room temperature. The cDNA fragments, which are bound to capped RNA transcripts, are enriched by Dynabeads M-270 (Invitrogen), and subsequently ligated to a double-stranded adaptor (5'-AGC TTC TAA CGA TGT ACG CTC GAG TCC AAC NN-3' and 3'-TCG AAG ATT GCT ACA TGC GAG CTC AGG TTGp-5') using T4 DNA ligase (NEB). For each candidate transcript to be validated, linker-ligated cDNAs were used as templates; PCR reaction was carried out with a junction primer (which spanning the linker and 5′ gene specific sequence of the TSS cluster mode) and a downstream gene-specific primer (100-200bp distance). As negative control, total RNA was pre-treated with TAP (Tobacco Acid Pyrophosphatase) and processed side-by-side with the RNA sample without TAP treatment (or with 5′ cap structure).

For the validation using oligo-capping strategy, total RNA of the fly embryo (0-24h) was BAP/TAP treated. A chimeric linker was ligated to the released 5′ phosphate group. Reverse transcription was performed following the same procedure as shown in the library construction. A junction primer spanning the linker and 5′ gene specific sequence of the TSS cluster mode, together with a downstream primer (100-200bp

distance), were used to carry out PCR reaction to validate 5′ sequence immediately downstream of the cap structure. The RNA sample without 5′-linker ligation was used as negative control.

6.3 Results

6.3.1 A Paired-End Strategy for Deep Sequencing of Cap-Trapped RNA

To accurately identify and investigate the process of transcription initiation in a complex eukaryotic genome, we developed the Paired-End Analysis of TSSs, or PEAT strategy (Figure 47a), which harnesses the paired-end capability of the Illumina/Solexa platform. Compared to the CAGE method (Carninci et al. 2005; Shiraki et al. 2003), our strategy has several advantages in experimental design. A circularization step is expected to improve the overall specificity of the library construction by removing undesired cDNA fragments without the matched adaptors; and the subsequent rolling circle amplification (RCA) can help reduce amplification biases and erroneous bases introduced during library construction (Esteban, Salas, and Blanco 1993; Hosono et al. 2003). More importantly, paired-end reads generated by the PEAT method can provide us with higher alignment yield and accuracy, as well as additional information on gene structure (e.g., linking 5′ TSS tags to known genes or transcripts).

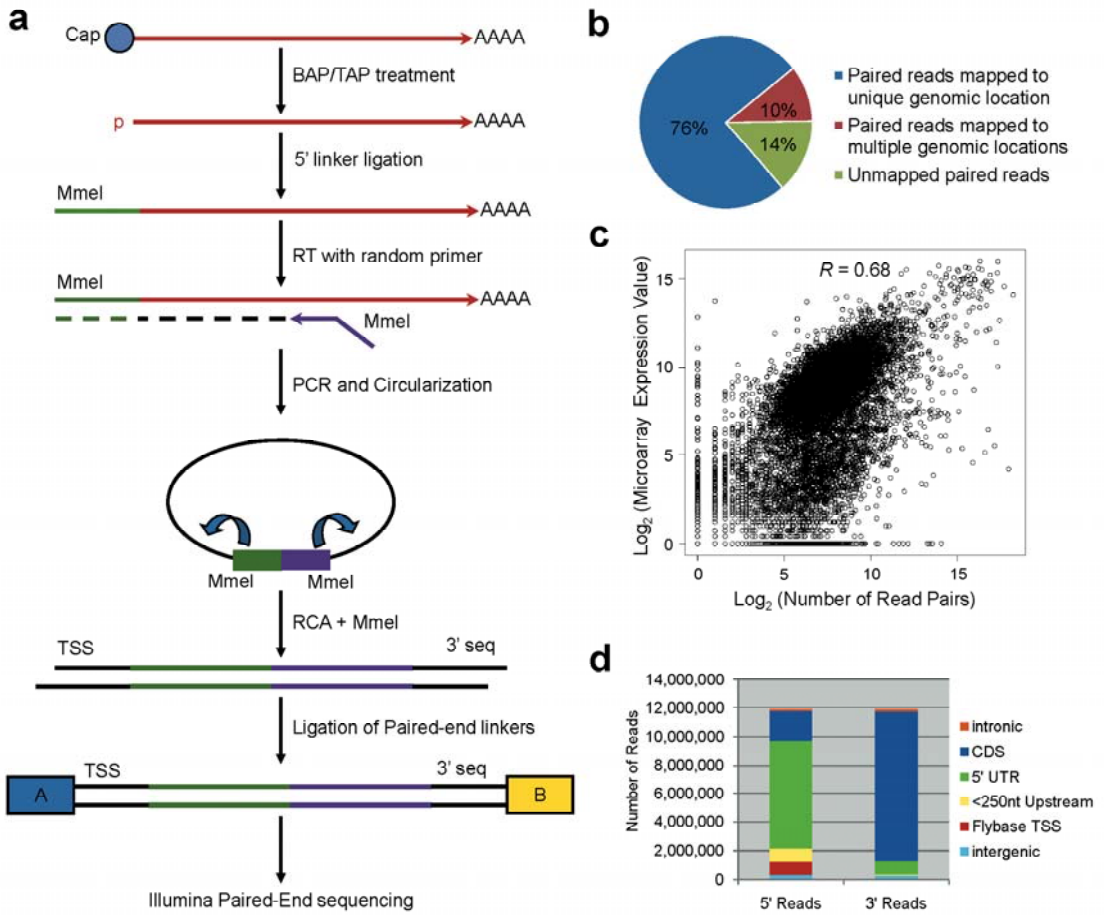


Figure 47: Paired-End Analysis of Transcriptional start sites (PEAT)

(a) Schematic outline of the PEAT strategy. The RNA fragment is shown as an arrowed line (red), the two Mme I sites induced at the oligo-capping and reverse transcription (RT) steps are shown in green and purple, respectively. Bridge ligation is used to circularize cDNA products. The resulting DNA circles are then RCA amplified and digested with Mme I to obtain ditags, each of which contain a TSS tag and a downstream 3′ tag linked by a generic sequence (green + purple). The Solexa adaptors (blue and yellow box) are ligated and the final PEAT library is sequenced by Illumina Genome Analyzer. **(b) Mapping efficiency of the reads that have built-in linker sequences, combined from two technical replicates. (c) Comparison between PEAT and microarray expression data.** 10,101 genes were plotted that had at least 1 mapped read-pair and were included in the microarray data. For the array data, expression level is the mean of simple background

subtraction values across 3 replicates from mixed stage 0-11 *D. melanogaster* embryos. To estimate the expression level using paired-end sequencing data, we used the counts of 3′ tags that map to a transcribed region. Correlation coefficient was determined by Pearson correlation. **(d) The distribution of uniquely mapped 5′ and 3′ reads relative to known TSSs and other genomic regions.**

Table 9: Summary of PEAT Generated Data

	Replicate 1	Replicate 2	Combined
Number of Read-Pairs with Identifiable Linker Sequences	8,258,735	7,470,183	15,728,918
Read-Pairs Mapped to a Unique Genomic Location	6,246,759	5,653,860	11,900,619
Read-Pairs Mapped to Multiple Genomic Locations	862,748	782,828	1,645,576
Non-Redundant Read-Pairs	4,103,558	3,752,136	7,062,714
Non-Redundant Read-Pairs Mapped to a Unique Genomic Location	1,688,228	1,569,274	2,716,981
Genes Represented by at Least 1 Read-Pair	11,111	11,073	11,418
Genes With An Identified Read Cluster Consisting of More Than 10 5′Reads	--	--	8,577
Genes With An Identified Read Cluster Consisting of More Than 50 5′Reads	--	--	5,563
Genes With An Identified Read Cluster Consisting of More Than 100 5′Reads	--	--	4,007

The PEAT strategy was employed to monitor global TSS usage in mixed stage embryos (0-24h) of *D. melanogaster.* We obtained 17.5 million raw paired-reads from two technical replicates. For approximately 90% of the paired-reads, both the TSS and 3′ reads were distinguishable by their built-in linker sequences (Table 9). Of those paired-reads, 76% were mapped to a unique location within Release 5 of the fly genome. An additional 10% of tag pairs mapped to multiple genomic locations (Figure 47b), possibly due to transposable elements or other regions with low sequence complexity (data not shown). The majority of 5′ reads were mapped to either a known TSS or its surrounding regions, confirming that our approach captured the very 5′ end of capped transcripts (Figure 48). The median distance between the 5′ and 3′ reads at the transcript level was 279nt (Figure 49) and the 3′ reads are mostly mapped to coding regions of annotated genes, indicating the success of the paired-end library construction.

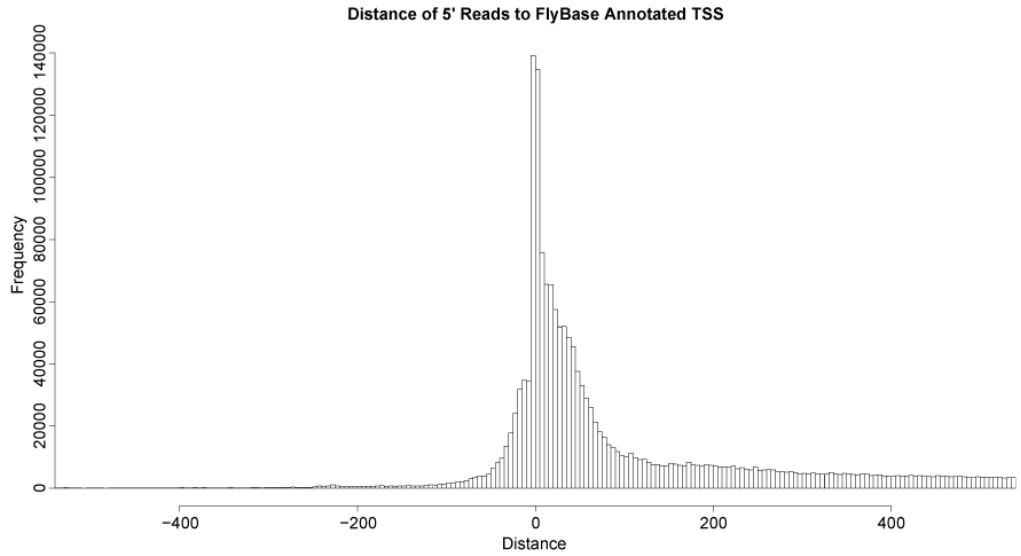


Figure 48: Distribution of 5′ reads relative to annotated FlyBase TSSs

All aligned 5′ reads were included in the analysis except those mapped to intronic regions; these reads were excluded because the distribution is based on the distance within the transcript from the TSS, not the overall genomic distance. For intergenic reads, the distance was determined based on the nearest downstream annotated TSS. The tag frequency for each group (binned for every 10 bp) is shown. The results showed a clear peak at the annotated TSSs with a long tail extending into regions downstream of TSSs. This is in agreement with previous observations in Chapter 3 (Rach et al. 2009) which indicated that Flybase annotations are generally based on the longest known transcript rather than the most frequent one (i.e. the mode of the cluster, as used in this study).

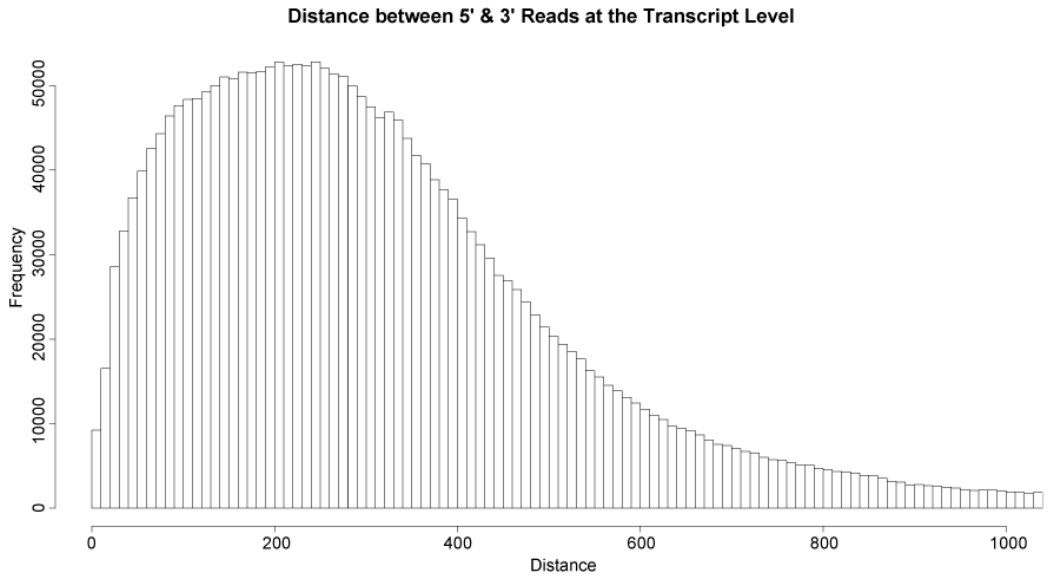


Figure 49: Distribution of the distance between 5′ and 3′ reads at the transcript level

The distance between 5′ and 3′ tags in each paired read was determined based on annotated transcripts in FlyBase. The distances were binned at a 10-bp resolution (X axis), and the abundance of each group is shown (Y axis). The median distance between the 5′ and 3′ reads at the transcript level was 279 bp.

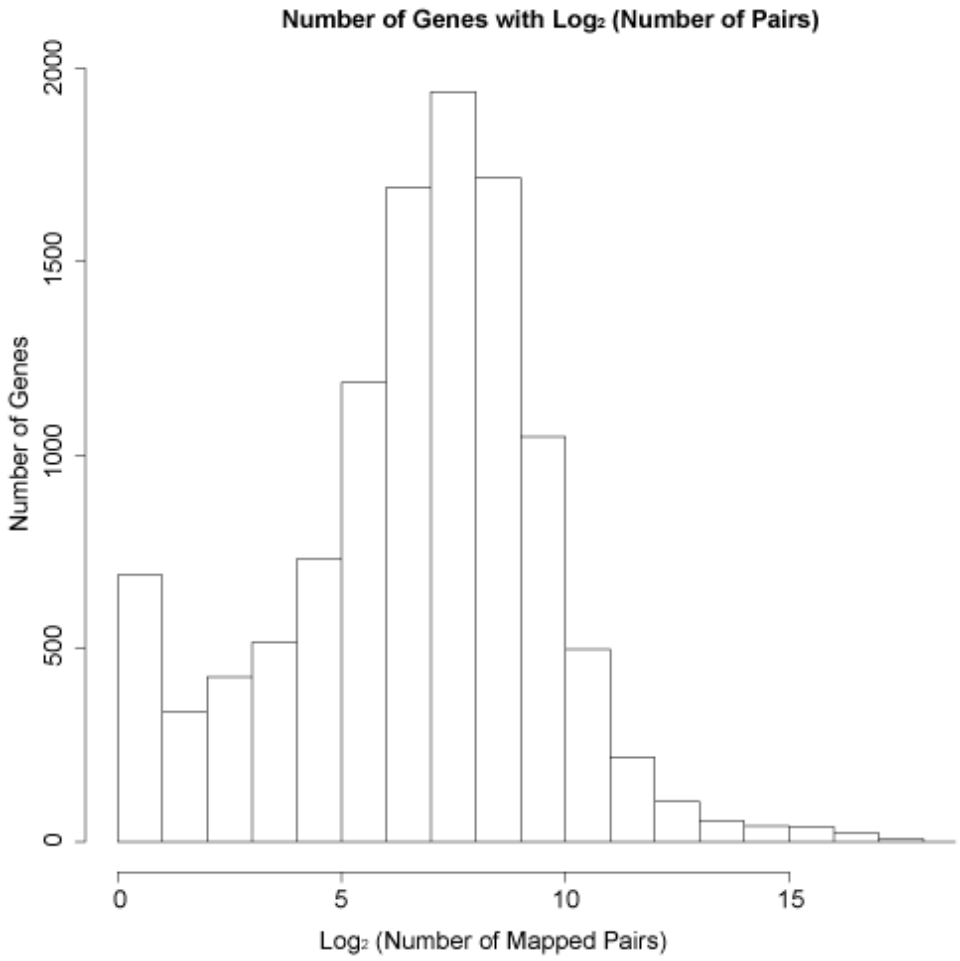


Figure 50: Distribution of TSS tag counts in annotated genes

The count of uniquely mapped 5′ TSSs in each gene was computed. The results were log2 transformed and plotted. On average, there are 256 tags per gene (log2 count =~ 8), indicating that our data set has deep coverage for annotated genes, including relatively rare transcripts.

The mapping results showed that on average there are ~256 paired tags per gene (Figure 50), suggesting that our data set has deep coverage for known genes. Notably, we found that 81.5% of genes currently annotated by FlyBase (v5.14) (Tweedie et al.

2009) were represented by at least one read-pair, consistent with the notion that eukaryotic genomes are broadly transcribed. Taken together, the mapping yield was considerably higher than that of deepCAGE (Valen et al. 2009). The fraction of aligned tags and coverage of the genome were also dramatically improved from a previous CAGE study of *D. melanogaster* (Ahsan et al. 2009) (Table 10, Figure 51).

Table 10: Comparison With a Previous *Drosophila* 5' CAGE Study

	PEAT*	**MachiBase****
Total Alignable Reads	15,728,918 (two channels)	5,619,701 (one channel)
Reads mapped to unique location	11,900,619	3,512,967
% of uniquely mapped reads	75.7%	62.5%
Reads mapped to multiple location	1,645,576	234,519
% of reads mapped to multiple locations	10.4%	4.2%
Non-redundant 5′ reads	1,168,474	306,829
Genes Represented by at Least 1 Read	11,418	10,196
Genes With An Identified Read Cluster Consisting of More Than 10 5′Reads***	8,577	4,799
Genes With An Identified Read Cluster Consisting of More Than 50 5′Reads***	5,563	2,406
Genes With An Identified Read Cluster Consisting of More Than 100 5′Reads***	4,007	1,644
% clusters within coding region containing ≥ 100 Reads	25%	19%
% clusters of class NP with ≥ 100 Reads	33%	37%
% clusters of class BP with ≥ 100 Reads	18%	23%

% clusters of class WP with ≥ 100 Reads	49%	40%

* Data of Embryo (0-24h)
** Data of Embryo as defined by Ahsan *et al.*(Ahsan et al. 2009)
*** Clusters within 250nt upstream of a currently annotated transcript were included

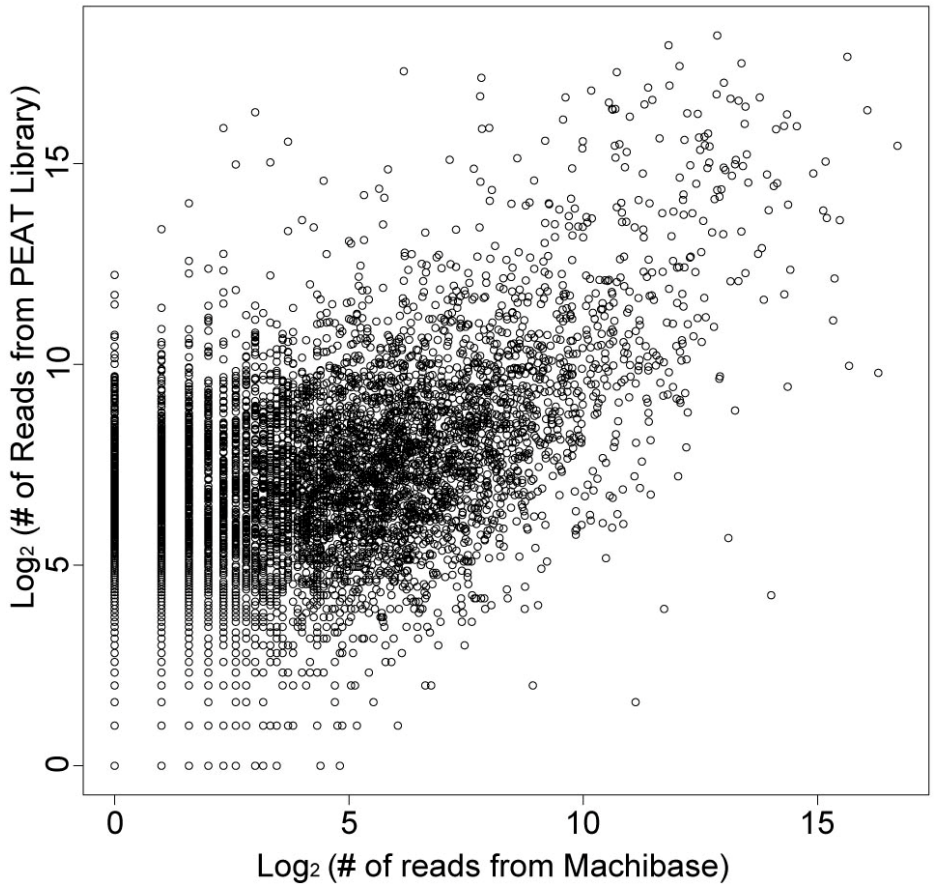


Figure 51: Comparison between the PEAT results and MachiBase

MachiBase is based on 5'CAGE tags only, and the , comparison of the read count per gene between the two datasets was therefore based on the number of 5' reads in the TSS proximal regions, including those that are ≤ 250nt upstream of an annotated TSS, overlap the TSS, or are found within the 5' UTR. The correlation coefficient (R = 0.53) between PEAT library and MachiBase is based on all genes

that had at least 1 tag in each dataset. The relatively low correlation might be in part due to the considerably larger number of reads per gene in the PEAT dataset compared to those available from MachiBase.

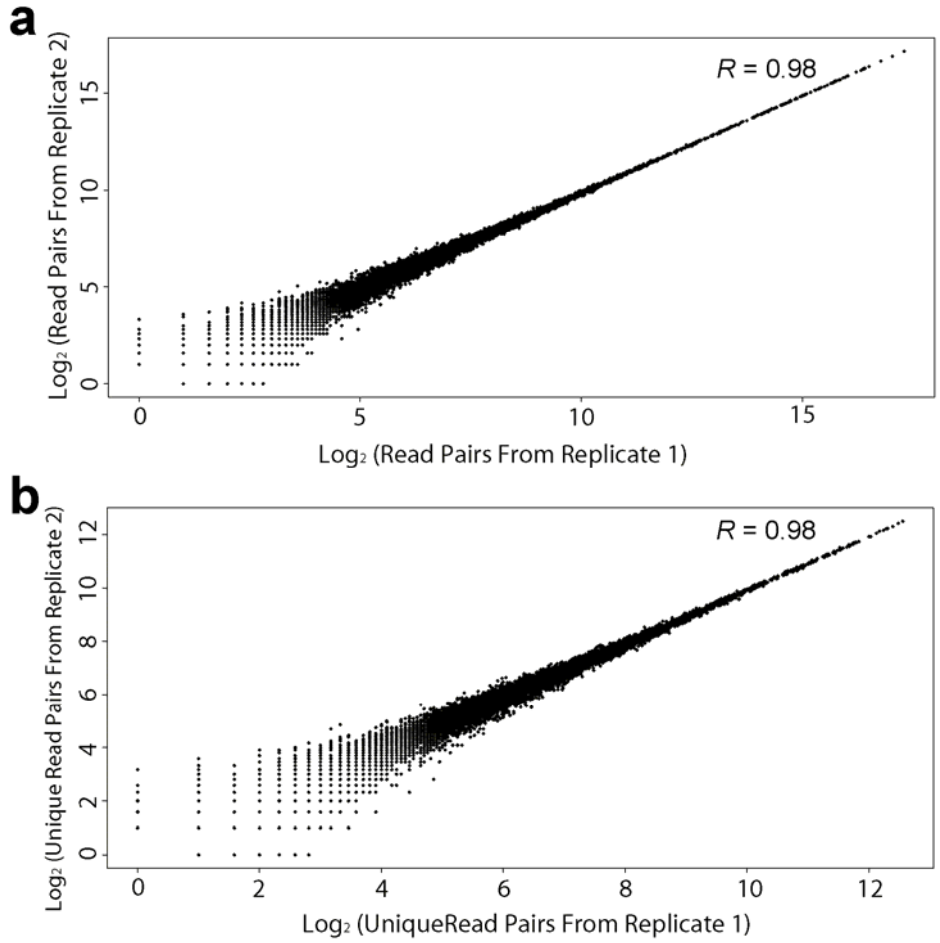


Figure 52: High reproducibility of the PEAT method

Correlation coefficient (R) between two replicates was computed based on the number of 3′ reads mapped to the transcribed region of individual genes. **(a)** Employs all read pairs mapped to a single gene. **(b)** The read pairs were further collapsed to non-redundant reads, i.e. multiple reads with identical 5′ and 3′ tags were counted once. Under both conditions, the two replicates showed very high correlation, suggesting the PEAT method is highly reproducible. The results also demonstrated that little bias was introduced during library construction since the non-redundant reads show consistent results as well. Thus, all mapped reads, rather than non-redundant reads, were used in the analyses of TSS clustering

and initiation pattern identification. This allowed us to improve the overall coverage and accuracy.

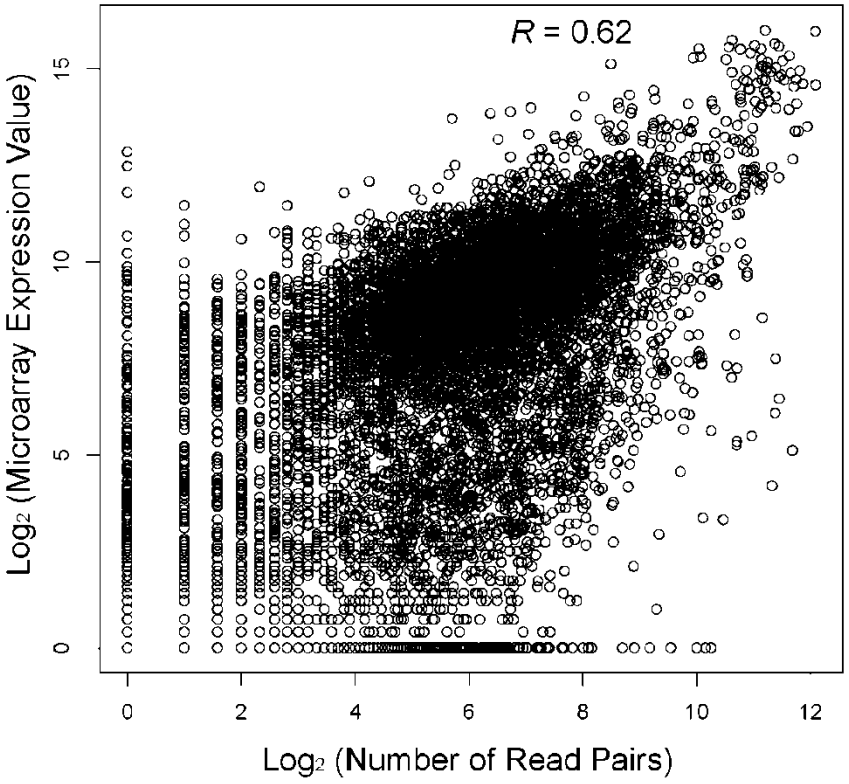


Figure 53: Comparison between the PEAT and microarray-based approach

Microarray dataset (Y axis) was obtained from GEO: GSE11880. The data set contained a mix of samples from *Drosophila* embryos of stages 0-11. Expression level of each gene was computed by averaging the signal minus background values from three separate arrays. To determine the expression level from the PEAT results (X axis), we computed the number of non-redundant read pairs by consolidating those with 3′ tags mapped in the same transcribed region. 10,101 genes were included in the analysis that were present in the array and had at least one read-pair mapped to them.

The two technical replicates were highly correlated (R = 0.98, Figure 52), indicating the reproducibility of PEAT. We next compared the PEAT results with a microarray-based expression dataset obtained from fly embryos of a similar broad developmental window (embryonic stages 0-11). With minimal normalization on both the array and sequence data, we observed that PEAT and array expression profiles are significantly correlated (R = 0.68, Figure 47c and Figure 53). The result is comparable to the correlation observed between microarray and standard RNA-Seq approaches (Wilhelm et al. 2008). Therefore, the read count of the PEAT method can potentially be used to estimate transcript abundance.

Notably, the paired-end strategy clearly allowed for a more accurate mapping of the short reads obtained. The addition of the 3′ reads enables ~4% of the 5′ reads to be aligned to a unique genomic location instead of multiple locations. Furthermore, the downstream tags can also correct assignment mistakes caused by sequencing errors. In fact, ~0.3% of the 5′ reads would have been wrongly aligned if the downstream tag had not been provided (Table 11). It is expected that such improvements will become more prominent for larger, e.g. mammalian genomes and/or when the sequencing error is relatively high (e.g. overcrowded Illumina/Solexa runs).

Table 11: Improved Mapping of Raw Data by Paired Reads

	Replicate 1	Replicate 2	Combined
Pairs with linker sequence	8,258,735	7,470,183	15,728,918
5′ reads that would have mapped to multiple locations	303,737	273,020	576,757
% of 5′ reads that would have mapped to multiple locations	3.68%	3.65%	3.67%
5′ reads that would have mapped to a different location	28,343	25,176	53,519
% of 5′ reads that would have mapped to different location	0.34%	0.34%	0.34%

6.3.2 Characterization of Read Clusters and Definition of Initiation Patterns

As another key parameter to evaluate the robustness of the PEAT method, we found that the majority of the 5′ reads are mapped to either a known TSS or its surrounding regions, confirming that the PEAT strategy captures the very 5′ end of capped transcripts (Figure 47d). Since high-throughput TSS maps have shown that mammalian promoters often exhibit diverse initiation patterns, one open question was whether *Drosophila* promoters would show a similar complexity of initiation patterns. To this end, we clustered the mapped 5′ reads (see Figure 54a, Methods). This resulted in 34,664 discrete clusters covering 8,577 genes, among which more than 5,500 genes have at least one cluster with ≥50 reads. While approximately half of the clusters overlap

annotated TSSs, a quarter of them are located within coding regions (Figure 54b). Such coding clusters have also been reported at similar frequency in mammals (for 12% of raw reads or 21.5% of TSS clusters) (Carninci et al. 2006; Fejes-Toth et al. 2009) and will be investigated in more detail below.

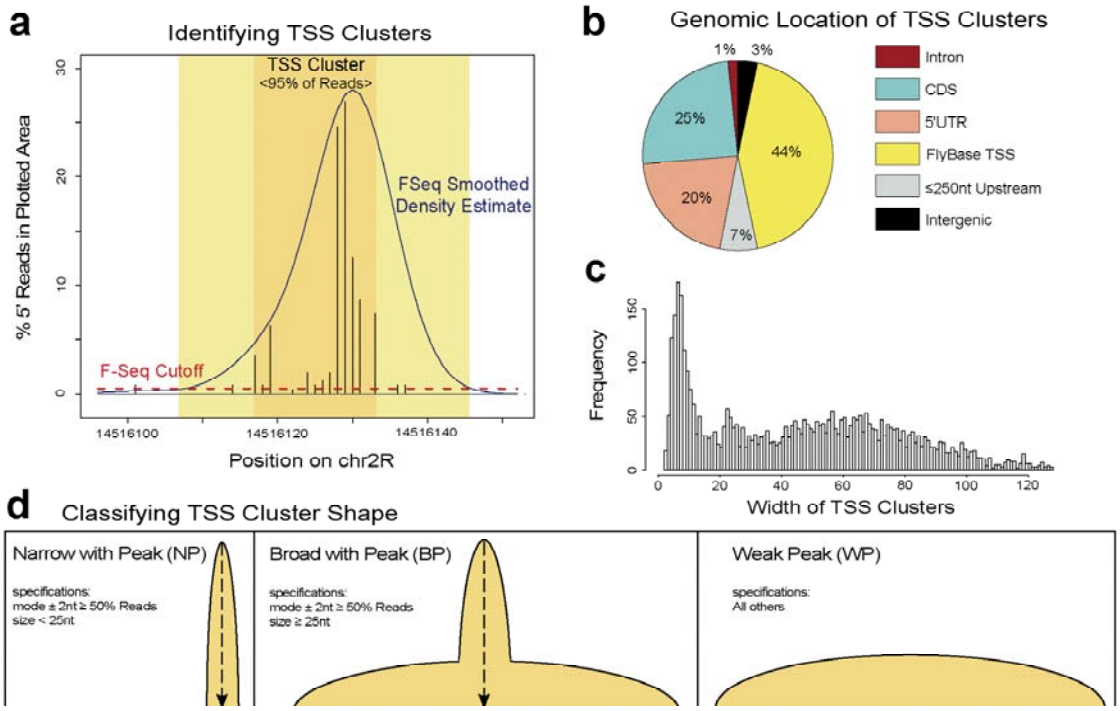


Figure 54: TSS clusters and initiation patterns identified in the *Drosophila* embryo.

(a) The approach for identifying TSS clusters. A representative example (chr2: 14516000-14516600) is shown. In essence, a smoothed density estimate of 5′ TSS tags was computed (blue line). Cluster boundary was then determined as exceeding a baseline score, estimated on a genomic background (red line). TSS clusters were further condensed to the shortest distance containing 95% of the reads (dark shaded area). **(b) The genomic locations of all clusters that contain ≥100 reads.** Clusters overlapping an annotated TSS in FlyBase were classified as FlyBase TSS. For the remaining clusters, classifications were based on the mode

of each given cluster and its relative location to annotated transcripts. **(c) Size distribution of all clusters with ≥100 reads.** Cluster sizes are similar to previous reports for mammals, with the majority of clusters shorter than 120nt in length. **(d) Definition of initiation patterns.**

In order to reliably identify/analyze the initiation patterns, we focused on 5' clusters with ≥ 100 reads (5,699 clusters in 4,007 genes). The cutoff was stringent to ensure high-quality assignments of initiation patterns and sequence motifs. The clusters spanned a broad size range, describing a complex multimodal distribution (Figure 54c) suggesting distinct initiation patterns. In fact, the cluster size distribution can be approximated by two Gaussian distributions, the intersection of which falls at a value of ~25nt. Read clusters were thus separated into three initiation patterns, Narrow with Peak (NP), Broad with Peak (BP) and Weak Peak (WP), along the two dimensions of cluster size and read distribution within each cluster (Figure 54d).

6.3.3 Initiation Patterns Are Linked to Specific Core Promoter Sequence Features

In mammals, 'peak' and 'broad' promoters tend to be associated with TATA box and CpG islands, respectively (Carninci et al. 2006). Since the fly genome does not contain CpG islands, it is intriguing to find that broad promoters do exist in *Drosophila*. We therefore aimed to determine whether distinct initiation patterns are associated with core promoter motifs previously defined in *Drosophila* (Ohler et al. 2002). We extracted 200 nt sequences centered on the mode of each cluster (i.e. the most frequent TSS within

the cluster). Promoter sequences were aligned for each initiation pattern and the results showed that initiation preferentially occurs at an adenine, immediately preceded by the 'TC' di-nucleotide for all 3 initiation patterns (Table 12 and Figure 55). The (T)CA consensus matched the minimal sequence requirements at the TSS as reported in other eukaryotes from yeast to mammals (Carninci et al. 2006; Zhang and Dietrich Mapping of transcription start sites in saccharomyces cerevisiae using 5' SAGE 2005), but is only a substring of the fly initiator motif as originally reported (Purnell, Emanuel, and Gilmour 1994). Thus, even for the broad pattern, defining the reference TSS at the mode is linked to a significant presence of a minimal initiator consensus.

Table 12: Frequency of Consensus di- and tri- Nucleotides Relative to the TSSs and Coding Region Clusters

Note: The +1 position within each cluster is defined by the mode of that cluster, i.e. oblivious to its location in the genome. We here show the analysis comparing coding region clusters to those near the start site of a gene. (Out of 5699 clusters, 426 clusters which fell into either intergenic or intronic regions were not included in the analysis).

	$T^{-2}C^{-1}A^{+1}$	$C^{-1}A^{+1}$	$T^{-1}C^{+1}A^{+2}$	$T^{-1}C^{+1}$
Narrow with Peak	550 (44%)	858 (68%)	13 (1%)	86 (7%)
Broad with Peak	274 (36%)	483 (64%)	13 (2%)	50 (7%)
Weak Peak	387 (19%)	973 (48%)	46 (2%)	128 (6%)
Coding Region Read Cluster	24 (2%)	108 (8%)	476 (35%)	804 (59%)

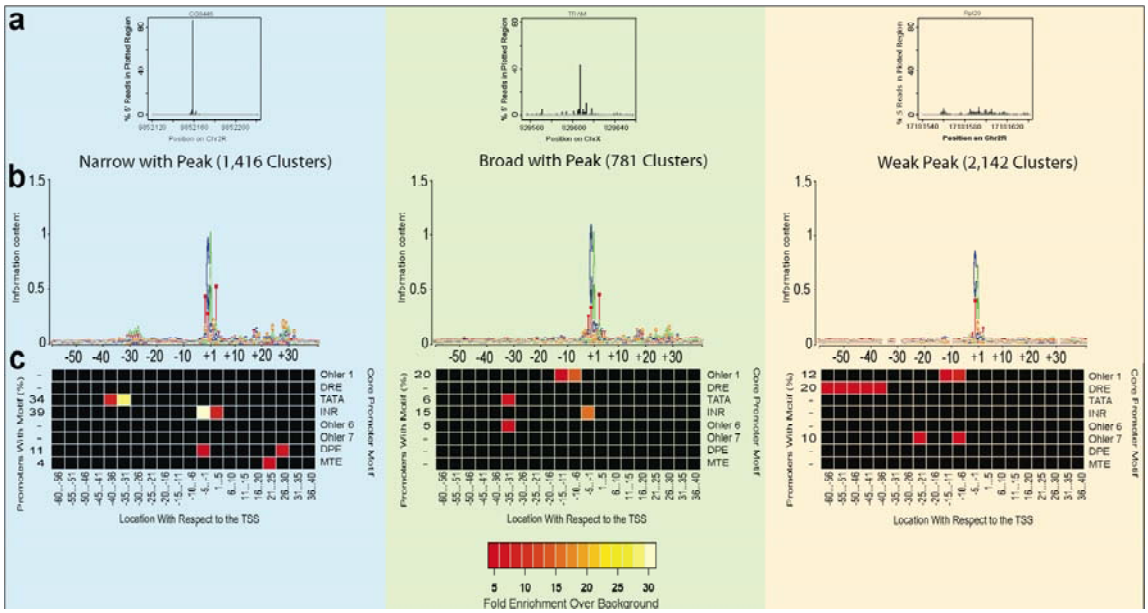


Figure 55: Promoter motifs associated with distinct promoter types.

(a) The three initiation patterns, NP, BP and WP, are each represented by a candidate locus. The graphs show the relative percentage of 5′ reads that are mapped within a 100nt window. **(b)** Sequence landscape in the promoter region of each pattern. The mode location of each cluster is set as reference point '+1'. Sequence logos (Schneider and Stephens 1990) of 100-nt window are shown. **(c)** The core promoter motifs overrepresented for each initiation pattern. Significant motifs were identified in 200nt core promoter sequences and binned into 5nt intervals; only the 100nt region surrounding the TSS is shown as no motifs were found to be enriched outside of this window. All bins with normalized motif occurrences of 5-fold enriched or above are shown. The percent of sequences containing at least one high-stringency instance of each motif in its preferred location is listed on the left side of the heat map. In NP promoters, the DPE was enriched at its known location (+26,+30) and at an additional site (-5,-1), which has previously been observed in mammalian data (Sandelin et al. 2007); the second location likely reflects some overlap in sequence similarity rather than functional DPE occurrences, as the importance of precise spacing has been clearly established (Burke and Kadonaga 1996; Kutach and Kadonaga 2000).

While the *Drosophila* genome does not contain CpG islands, a variety of sequence motifs have been reported to be present within the core promoter region (Juven-Gershon et al. 2008; Ohler and Wassarman 2010). Included in this set of motifs are the location-specific TATA box and INR motif, both of which have been well characterized throughout many eukaryotic genomes. In addition, the Motif Ten Element (MTE), the Downstream Promoter Element (DPE) (Kutach and Kadonaga 2000), and the DNA-replication related element (DRE) (Hochheimer et al. 2002) have all been experimentally validated as functional core promoter motifs in fly. Computational analysis has led to the identification of additional motifs (Ohler et al. 2002), three of which (Ohler1, 6, & 7) have consistently been reported in multiple studies since their initial identification (FitzGerald et al. 2006; Sharan and Myers 2005).

We evaluated the presence and preferred locations of these eight motifs in different initiation patterns by scanning for stringent matches to positional weight matrices (see Appendix G). Strikingly, the results revealed distinct associations between initiation patterns and sequence motifs. The canonical core promoter motifs with previously known location bias (TATA, INR, DPE, MTE) were highly associated with NP promoters (Figure 55a), while WP promoters were strongly associated with 3 motifs (Ohler1, DRE, Ohler7) and showed a moderate enrichment for Ohler6 (Figure 55b, 55c and Figure 56). BP promoters, which have characteristics of both NP and WP promoters, showed a combination of the most frequent motifs found in both the NP and WP

promoters. The largest span of motif enrichment was 25 nt for the DRE motif in WP promoters, reflecting the broad initiation pattern in this class. We noted that the INR and Ohler1 motifs share a strong conserved 'TCA' tri-nucleotide, i.e. the minimal initiator consensus described above. Likewise, Ohler6 is enriched at the same location as the TATA box and contains a minimal TAT consensus that is shared with the canonical TATA motif. Thus, it is plausible that Ohler6/Ohler1 is an alternative to the classic TATA/INR motif pair, an observation that has only become apparent with high-resolution data generated by PEAT. These results strongly argue that the initiation patterns in fly directly reflect the presence of the specific core promoter motifs that define them.

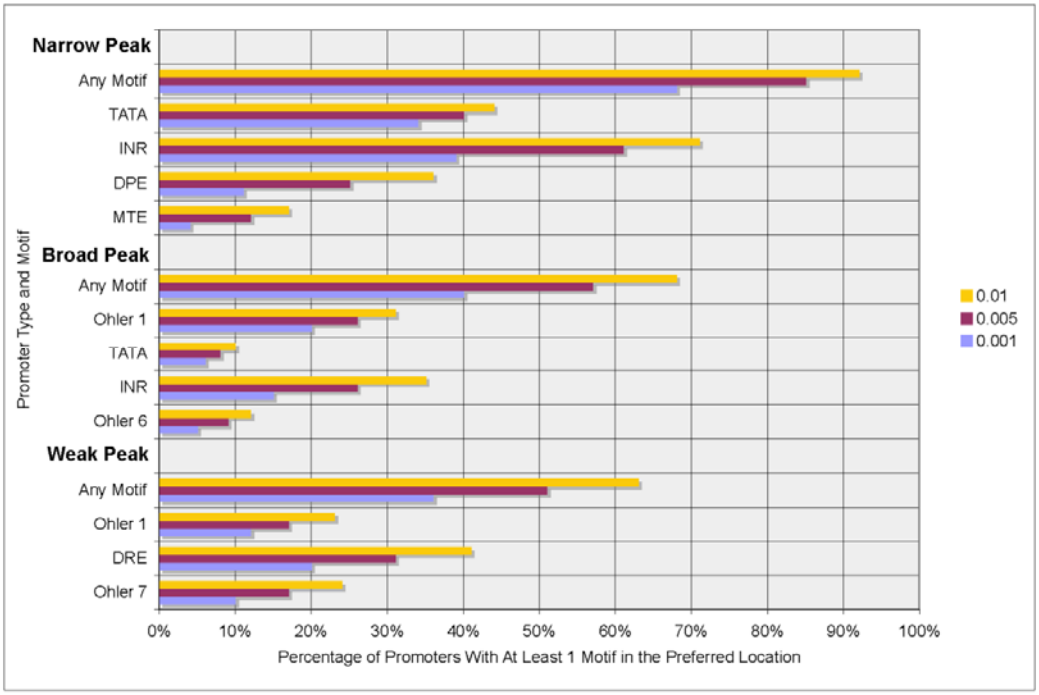


Figure 56: Motif prevalence at preferred locations

Motif searches were carried out with PATSER in the promoter sequences for each initiation pattern. The analyses were conducted at different P-value stringency thresholds: 0.001, 0.005 and 0.01. For each known promoter motif, the results are presented as the percentage of the promoters with at least one respective motif found at the preferred location. As expected, the motif prevalence increases as the threshold stringency decreases. Results at the most stringent condition (0.001) are shown in Figure 55.

6.3.4 TBP and TRF2 Binding Profiles Distinguish Different Initiation Patterns and Support Internal Re-capping

Analyses of core promoter sequences in *Drosophila melanogaster* have identified a set of overrepresented motifs (FitzGerald et al. 2006; Ohler et al. 2002). These elements include the canonical position specific motifs shared among metazoans, including the TATA box, which is found ~30nt upstream of the TSS, the initiator element (INR), which is located directly at the point of transcription initiation, and motifs downstream of the TSS, including the downstream promoter element (DPE). TATA boxes are recognized by the TATA-box binding protein (TBP), a component of the basal factor TFIID. Other enriched motifs are fly specific and include the DRE motif, which has less positional bias than the canonical elements, and is associated with core promoters bound by the TBP-related factor 2 (TRF2), which substitutes for TBP in remodeled basal complexes (Rabenstein et al. 1999).

Previous studies have therefore demonstrated that alternative components of Pol II complexes in *Drosophila* have different binding preferences and functions (Butler and Kadonaga 2001; Holmes and Tjian 2000; Isogai et al. 2007; Lee et al. 2005). Given the

differences we observe in initiation patterns and the varied motif occurrences between them, we further investigated whether distinct binding factors might be associated with the three initiation patterns defined here. As mentioned above, TBP and TRF2 are two transcription factors known to be components of different basal complexes in *Drosophila* (Hochheimer et al. 2002). We assessed the binding of these factors using previously published ChIP-chip data obtained from *Drosophila* S2 embryonic cell lines (Isogai et al. 2007), which provided sets of regions to which either TBP, TRF2, or a combination of both was bound. Due to the limited resolution of ChIP-chip assays, a subset of 1,138 NP, 660 BP, and 1,849 WP clusters separated by 500nt from any other cluster were evaluated.

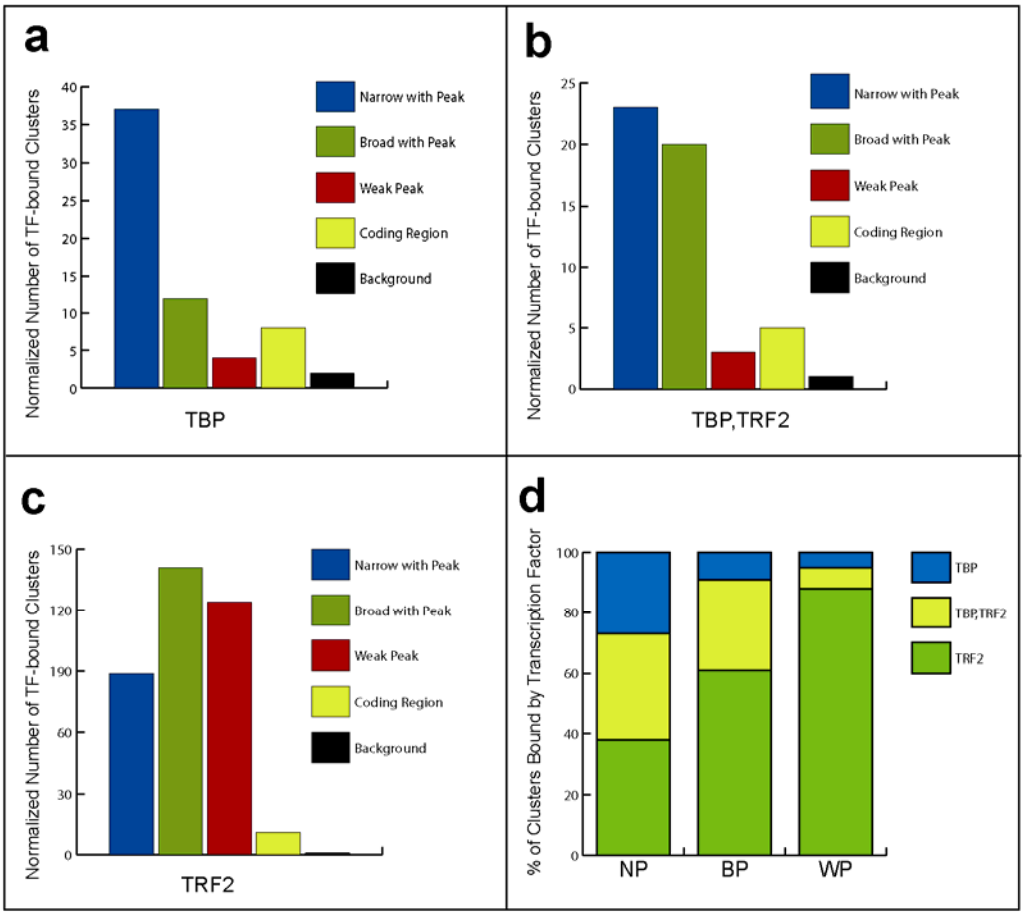


Figure 57: TBP and TRF2 are associated with different promoter classes and/or core promoter motifs

(a-c) Number of clusters overlapping binding sites for TBP, TRF2, or both TBP and TRF2 were counted and normalized to the number of occurrences per 1,000 read clusters. **(d)** Relative frequencies of factors bound to the different shape classes, for those clusters that have at least one of the factors bound.

Given the different experimental conditions, we did not expect perfect agreement between the ChIP-chip and our TSS data. Yet, we found that NP promoters had over 3-fold the amount of TBP and TBP&TRF2 binding to them than WP promoters (Figure 57a-b), while WP promoters had greater levels of TRF2 binding than NP

promoters (Figure 57c). BP promoters showed TF binding patterns similar to both NP and WP promoter types, as they have an intermediate level of TBP binding, a high amount of TRF2 binding, and high levels of both TFs binding. This further supports that these promoters are a potential hybrid of both the NP and WP classes, which agrees with both the definition of the initiation pattern and their motif composition. Across all initiation patterns, the ChIP-chip data showed higher numbers of promoters to be bound by TRF2 than by TBP. In agreement with potential experiment-wide differences, higher occupancy of Pol II was observed for regions bound by TRF2 than bound by TBP (Figure 58). As such, the associations between NP promoters with TBP binding and WP promoters with TRF2 binding were clearer when the relative frequency of each ChIP was compared separately among the clusters within each initiation pattern. NP promoters had the highest percentage of TBP binding (27%), followed by BP promoters (9%) and WP promoters (5%) (Figure 57d). Conversely, a higher frequency of TRF2 binding was observed in WP promoters (88%) followed by BP promoters (61%) and then NP promoters (38%). Differences in TBP binding also correlated with the frequency of a TATA box, which is the target sequence motif of TBP (Table 13).

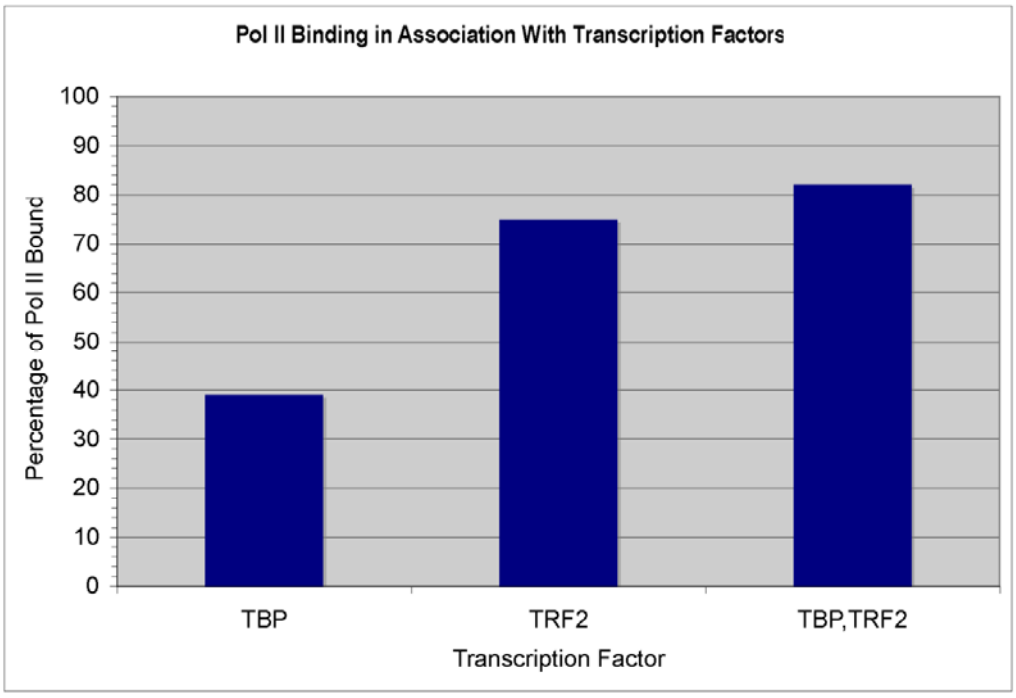


Figure 58: Pol II binding in association with transcription factors

The percentage of Pol II bound sites was calculated by dividing the number of sites bound by Pol II to the total number of sites bound by each transcription factor (TF).

Table 13: Frequency of TSS Clusters Bound By TBP, TRF2, or Both in Promoters With or Without the TATA-box

TSS Pattern	TF	TATA	No TATA
Narrow with Peak	TBP	76%	24%
	TRF2	11%	89%
	TBP,TRF2	0%	100%
Broad with Peak	TBP	63%	37%
	TRF2	19%	81%
	TBP,TRF2	15%	85%
Weak Peak	TBP	37%	63%
	TRF2	13%	87%
	TBP,TRF2	0%	100%

Taken together, NP promoters show an enrichment of TATA boxes, and the functional relevance of this is confirmed by the preferential binding of TBP, a subunit of the TFIID transcription initiation complex which is known to directly bind to the TATA box motif (Dynlacht, Hoey, and Tjian 1991). In contrast, the WP promoters, which lack bias towards a specific location for initiation, were found to be enriched for the DRE motif, which is preferentially bound by TRF2-associated with DREF, subcomponents of

an alternative initiation complex (Hochheimer et al. 2002). The presence of WP promoters in *Drosophila* is an interesting observation by itself as broad promoters in mammals are enriched for CpG islands (Carninci et al. 2006), a genomic feature not present in the fly. However, mammalian studies have so far not uncovered that different basal complexes bind to peaked and broad promoters; our findings here suggest the possibility that distinct complexes, similar to TRF2, may be associated with CpG islands in mammals.

Finally, binding of TBP and/or TRF2 was strongly under-represented in internally capped clusters (Figure 57a-c), providing additional support to the motif analysis, which had shown that the surrounding sequences were depleted of the known *Drosophila* core promoter elements.

6.3.5 Identification of Novel Transcription Start Sites

Paired-reads can also facilitate the direct link of novel TSSs to their respective genes. This becomes particularly important when TSSs are located distal from an annotated transcript. Previous technologies required the assumption that a potential novel TSS belongs to the nearest gene (Carninci et al. 2006), or had to rely on expression correlation between TSSs and the downstream gene to which they may belong (Manak et al. 2006). Because the median distance between the 5′ and 3′ tags is 279nt, identification of distal TSS by the PEAT method is not expected to be exhaustive. Of

342,943 read pairs where the 5′ read fell more than 250 nt upstream of an annotated TSS, 58,415 had the corresponding 3′ read mapped to the transcribed region of the downstream gene. From the clusters defined by these read-pairs, we selected 10 novel TSSs, meeting stringent criteria for experimental validation (Table 14).

Table 14: Candidate Distal TSSs Selected for Validation

Gene	Strand	Chromosome	TSS Location	Tested	Confirmed by oligo-capping	Confirmed by cap-trapping	TSS Cluster Shape
FBgn0028537	-	2L	14741837	Yes	+	+	NP
FBgn0033113	+	2R	2768500	Yes	-	-	NP
FBgn0033068	-	2R	2083769*	Yes	+	+	NP
FBgn0033688	-	2R	8050517	Yes	+	+	BP
FBgn0260442	+	3L	8553290	Yes	-	+	NP
FBgn0004228	-	3L	15512716	Yes	-	+	NP
FBgn0037410	+	3R	2035098	Yes	+	+	NP
FBgn0085320	-	3R	22251501	Yes	+	+	NP
FBgn0030136	+	X	9319746	Yes	+	-	NP
FBgn0024366	-	X	1232255	Yes	+	+	NP

*Also Identified by Manak *et al.*(Manak et al. 2006)

Two independent approaches (oligo-capping and cap-trapping) were used; 7 and 8 out of 10 candidates were validated, respectively (Figures 59 and 60). Since the reverse primers were designed to be within 100-200 nt of the TSS, the actual positive rate could be even higher considering that some of the failed sites might be caused by a splicing event downstream of the TSSs. Despite the high validation rate, we did not expect a substantial number of novel first exons for known protein-coding genes, as had been

proposed by a previous tiling array study (Manak et al. 2006). Distal TSS tags were rare in general (0.5% of the total reads) and the majority of them did not form clusters. Indeed, comparison with the tiling array study only resulted in 57 clusters with ≥10 reads being identified near one of their novel exons. Only 8 of these 57 clusters contained more than 100 reads, 1 of which was included in the experimentally validated set. One possibility is that most of the distal TSSs found in tiling arrays may be the result of non-polyadenylated RNAs, which would not be represented in our dataset due to poly(A)+ selection.

Order of target genes (FlyBase) in Figures 59 and 60:

1: FBgn0028537	2: FBgn0033068	3: FBgn0037410	4: FBgn0033113
5: FBgn0033688	6: FBgn0260442	7: FBgn0004228	8: FBgn0085320
9: FBgn0030136	10: FBgn0024366		

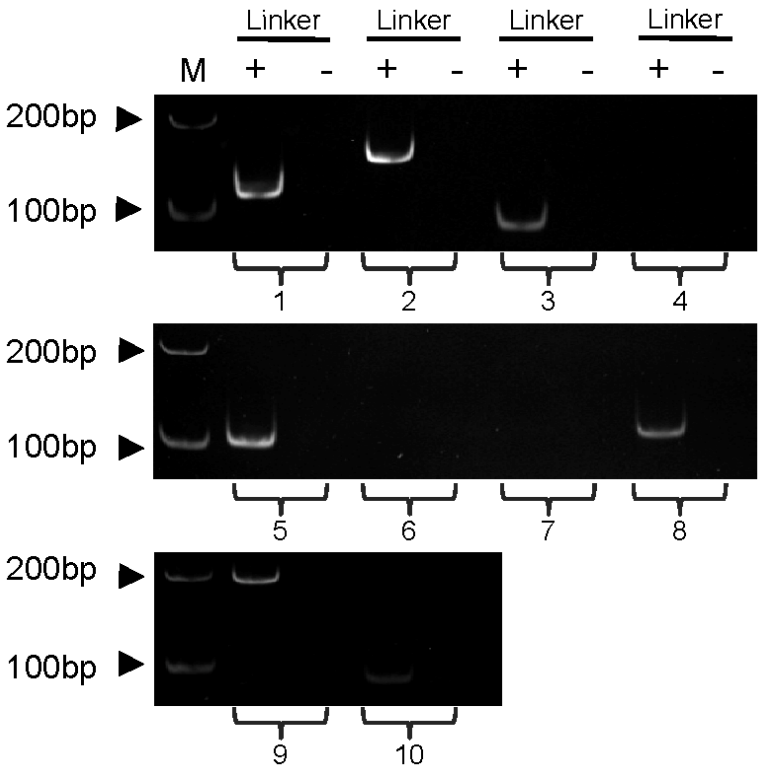


Figure 59: Validation of novel TSSs by oligo-capping

Ten novel TSSs were selected for validation by RT-PCR assay. To ensure that the final amplification products are derived from capped transcripts, junction primers were used, each of which contains a partial 5′ linker sequence (resulting from oligo-capping) and a gene-specific sequence at the mode of a putative TSS. The reverse primer was designed to be 100-200bp downstream of the candidate novel TSS. Amplification specificity was controlled by comparing cDNA fragments generated from RNA samples with (+ linker) and without (- linker) oligo-capping. Specific bands with the expected size were detected in 7 out the 10 cases, confirming novel TSSs at the distant sites.

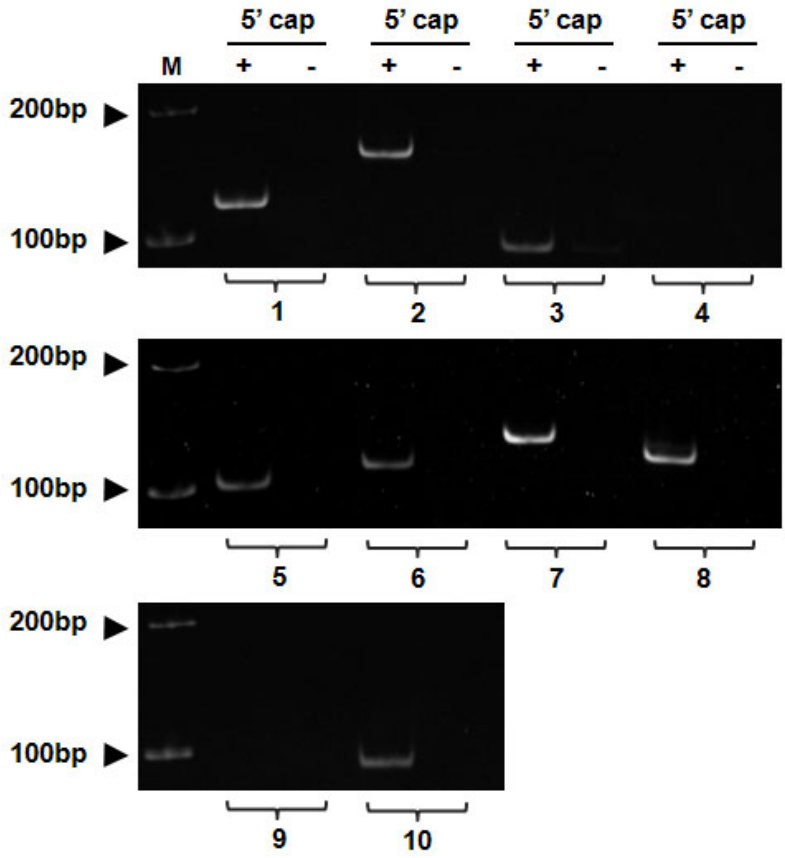


Figure 60: Validation of novel TSSs by cap-trapping

As an independent approach, cap-trapping was used to validate the 10 novel TSSs shown in Figure 59. As negative control, total RNAs were pretreated with TAP to remove the cap. The resulting RNAs (- 5′cap) were processed side-by-side with the RNAs with 5′ cap (+ 5′ cap) along the cap-trapping procedure (see Supplementary Methods for detail). To ensure that the final amplification products are derived from capped transcripts, junction primers were used, each of which contains a partial 5′ linker sequence (resulting from cap-trapping) and a gene-specific sequence at the mode of a putative TSS. The reverse primer was designed to be 100-200bp downstream of the candidate novel TSS. Specific bands with the expected size were detected in 8 out the 10 cases, confirming novel TSSs at the distant sites.

6.3.6 5' Capped Read Clusters in Coding Regions

In the initial clustering of reads, we observed that 25% of called clusters were found within the coding region of an annotated gene, and cluster analysis showed that the majority of them belong to the WP class (Table 15). Twelve candidates were selected for validation, and 10 were confirmed by two independent methods (83.3% of the cases; Table 16, Figure 61, Figures 62-64). Therefore, these clusters were not artifacts of the high-throughput protocol and indeed contained a 5' cap. Supporting this notion, recent studies in mammals have also identified a high prevalence of capped transcripts originated from the coding regions (Carninci et al. 2006; Fejes-Toth et al. 2009).

Table 15: Three Classes of Capped Clusters in Coding Region

	Narrow with Peak (NP)	**Broad with Peak (BP)**	**Weak Peak (WP)**
# of clusters	73	125	1,162
# of clusters with at least 1 TFBS in a preferred location*	2	3	67

* Preferred binding location for any given transcription factor can be seen in Figure 55. A *p*-value cutoff of 0.001 was used for the motif search.

Table 16: Candidate Internal Capping Selected for Validation

Gene	Strand	Chromosome	Cluster Mode Location	Confirmed by oligo-capping	Confirmed by cap-trapping
FBgn0003870	+	3R	27553415	+	+
FBgn0014269	-	2R	14858076	+	+
FBgn0020443	+	2L	12438946	+	+
FBgn0024841	+	3R	25549175	+	+
FBgn0026188	+	3R	5271105	-	-
FBgn0029629	-	X	2503629	+	+
FBgn0030341	-	X	11746055	+	+
FBgn0031769	+	2L	6052196	+	+
FBgn0035121	-	3L	259680	+	+
FBgn0037301	+	3R	1058801	+	+
FBgn0037707	+	3R	5379226	+	+
FBgn0051729	+	2L	13299058	+	-

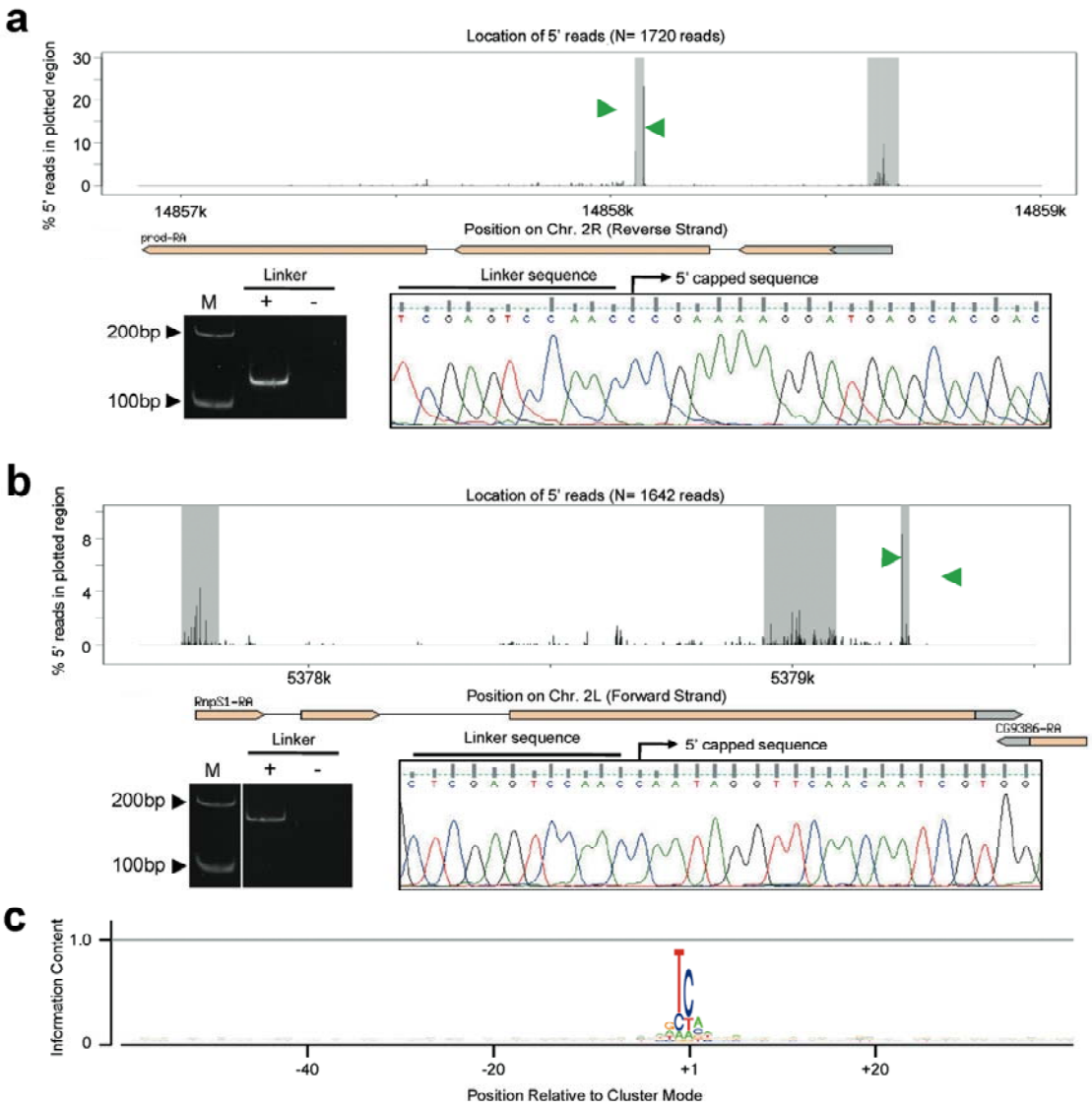


Figure 61: A distinct sequence motif identified for internally capped transcripts.

(a-b) The gene structures of the PROD and RNPS1 loci indicating exons (thick bar) and introns (thin bar) from FlyBase are shown. A thick grey bar represents the UTR region. Grey areas highlight read clusters (≥100 reads/cluster). Green arrows denote primer locations for RT-PCR validation. A junction primer, which spans the linker and 5′ gene specific sequence at the cluster mode, together with a downstream primer (100-200bp distance) were used to carry out RT-PCR. For

each locus, cDNAs derived from RNA samples with (+) or without (-) linker ligation were used as template. The DNA ladder (M) is shown in the left lane. Sanger sequencing results show the correct position of the mode of the called TSS cluster for (a) a capped 5′ read cluster in the middle of a coding region; and (b) an example of a capped 5′ read cluster near the end of the coding region. **(c)** Sequence logo(Schneider and Stephens 1990) of a 100nt window around the mode location (identified as '+1') of all clusters containing more than 100 reads and mapping to a coding region.

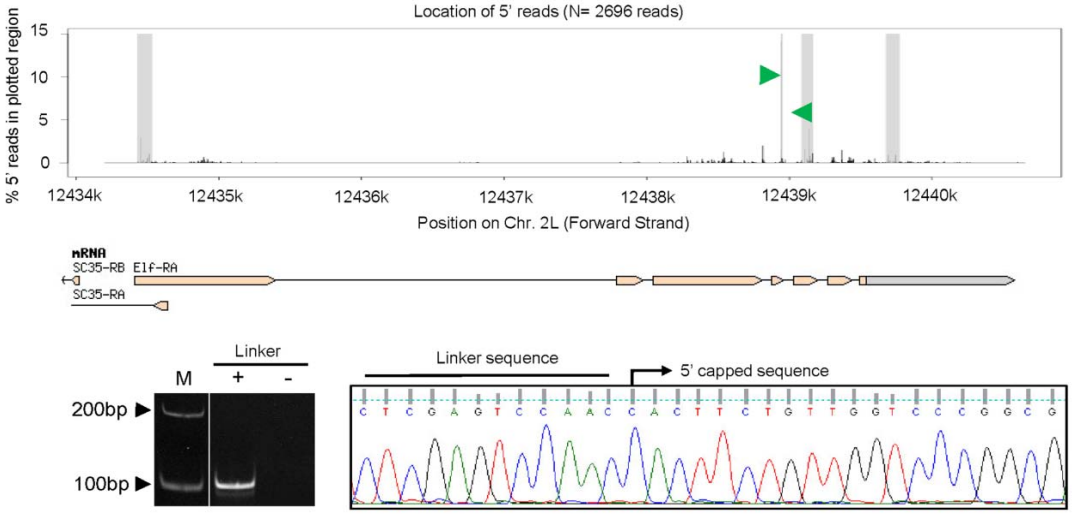


Figure 62: A validated example of internally capped transcripts

Analogous to Figure 61, the 5′ TSS reads obtained from PEAT and the FlyBase annotation are shown for the another example, the Elf gene locus (upper panel). The areas shaded in grey represents the called read clusters (≥100 reads/cluster). Two gene-specific primers (green arrows) were designed; the forward primer contains a partial sequence of a 5′ linker, which was added to all capped transcripts by oligo-capping. PCR amplification was conducted with first-strand cDNA generated from total RNAs (0-24hr embryos) with or without oligo-capping (+ or – linker). The latter served as a negative control to ensure all amplification products are generated from capped transcripts. Correct amplification products were detected as resolved by gel electrophoresis (lower left panel). The PCR products were further confirmed by Sanger sequencing and the junction region (5′ linker and its immediate downstream sequence) is shown.

Candidate genes selected for validation in Figures 63 and 64:

1: FBgn0014269	2: FBgn0020443	3: FBgn0029629	4: FBgn0030341
5: FBgn0031769	6: FBgn0035121	7: FBgn0037301	8: FBgn0037707
9: FBgn0003870	10: FBgn0024841	11: FBgn0026188	12: FBgn0051729

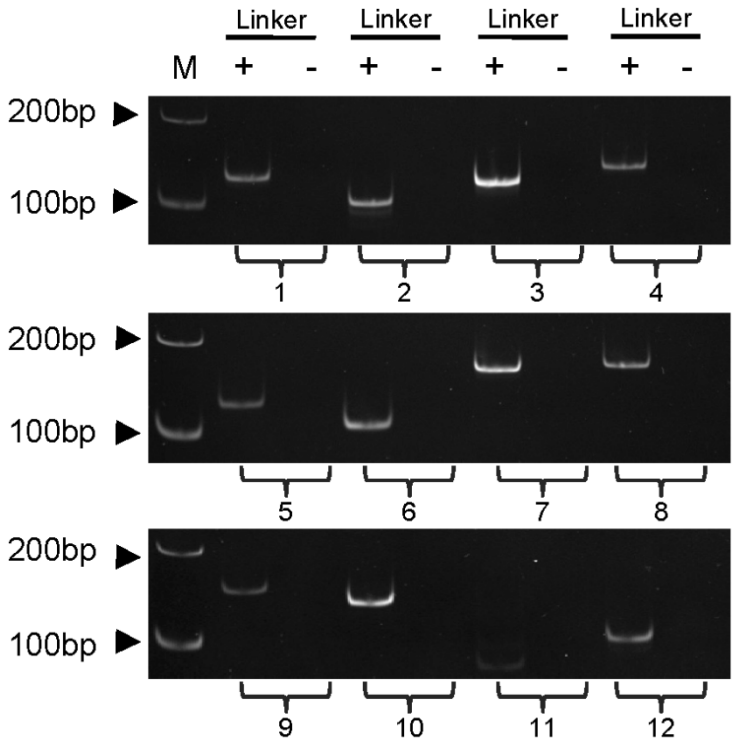


Figure 63: Validation of 5′ capped reads in CDS by oligo-capping

12 cases of internally capped transcripts were selected for further validation. RT-PCR was performed and a similar strategy was used as shown in Figure 61 and Figure 62. For each primer pair, the distance between two gene-specific primers was between 100-200 bp. Two RT-PCR reactions were performed for each locus using RNA samples with or without oligo-capping. The PCR products were resolved by gel electrophoresis. Correct bands with expected sizes were detected in 11 out of the 12 cases. For the single failed case (sample 11), a weak band was also observed (< 100 bp) and is likely due to nonspecific amplification.

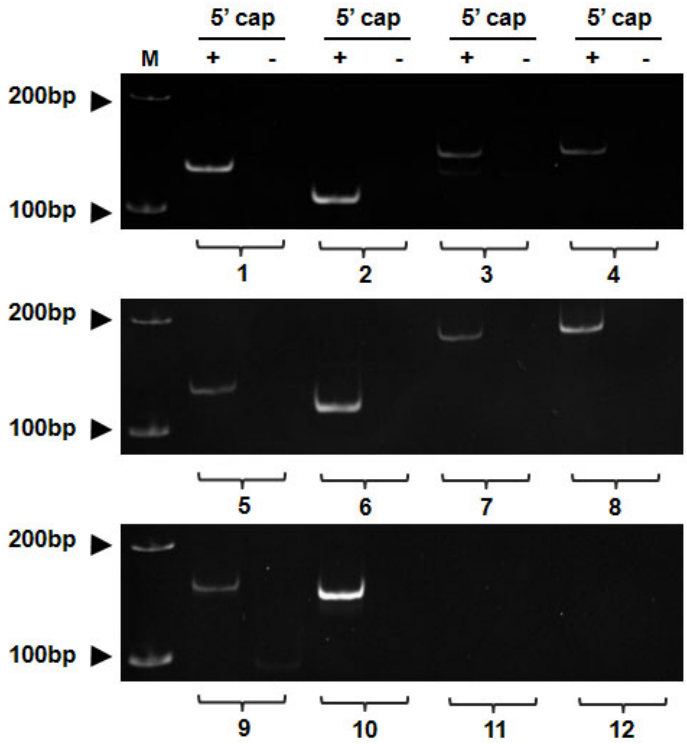


Figure 64: Validation of 5′ capped reads in CDS by cap-trapping

As an independent approach, cap-trapping was used again to validate the 12 cases of internally capped transcripts shown in Figure 63. The same experimental procedure as described in Figure 60 was used. Correct bands with expected sizes were detected in 10 out of the 12 cases. 10 candidates were validated by both oligo-capping and cap-trapping methods.

Several mechanisms may underlie the biogenesis of internally capped transcripts. First, they might result from *bona fide* alternative start sites in the coding region. Alternatively, these transcripts may be derived from longer precursors, for which the internal cap is introduced posttranscriptionally by a recapping mechaism (Fejes-Toth et al. 2009). Multiple lines of evidence from our data support the latter

model. First, searching the 200nt sequences surrounding the coding clusters revealed no overrepresentation of any of the core promoter motifs observed near bona fide TSSs (see Table 15); this is in agreement with our previous observation of a lack of promoter motifs around mammalian coding clusters (Megraw et al. 2009). The analysis of ChIP data (Isogai et al. 2007) showed frequent binding of TFs (TBP and/or TRF2) at TSS clusters but not at coding clusters. Together, our data suggests that 5′ capped coding clusters are unlikely initiated by Pol II.

In addition, we found that for 69% of the loci having a called read cluster within the coding region, a larger cluster (with more reads) was identified near the annotated TSS (Figure 65), indicating that internally capped transcripts are often accompanied by more abundant full-length transcripts. Moreover, the locations of the 5′ coding region read clusters spread evenly across the exons except for a lack of clusters at the far most 3′ end of the exon (Figure 66), similar to what has been reported in mammals (Fejes-Toth et al. 2009). The earlier study relied on TSS reads mapped across exon junctions, which are a tiny fraction of the total reads, to argue that recapping is a posttranscriptional event. Unique to the PEAT data set, we observed that the downstream paired tags of the coding clusters are predominantly located in the well-annotated exons rather than introns (~100-fold enrichment). Our results clearly demonstrate that internally capped transcripts are spliced or at least partially spliced, and the information on local transcript structure might be valuable for further characterization of the recapping mechanism.

Interestingly, we observed a distinct short sequence motif when aligning the sequences surrounding the mode of coding clusters. While this motif is at first glance reminiscent of the minimal initiator motif found in TSS clusters, it exhibits unique properties. 'CA' is the most frequent di-nucleotide at the -1 position in TSS read clusters, while the most prominent di-nucleotide at the mode location within coding region clusters is 'TC' (Table 12). Although the molecular mechanism of recapping remains elusive, the distinct motif implies that recapping might depend on specific sequences and/or protein factors.

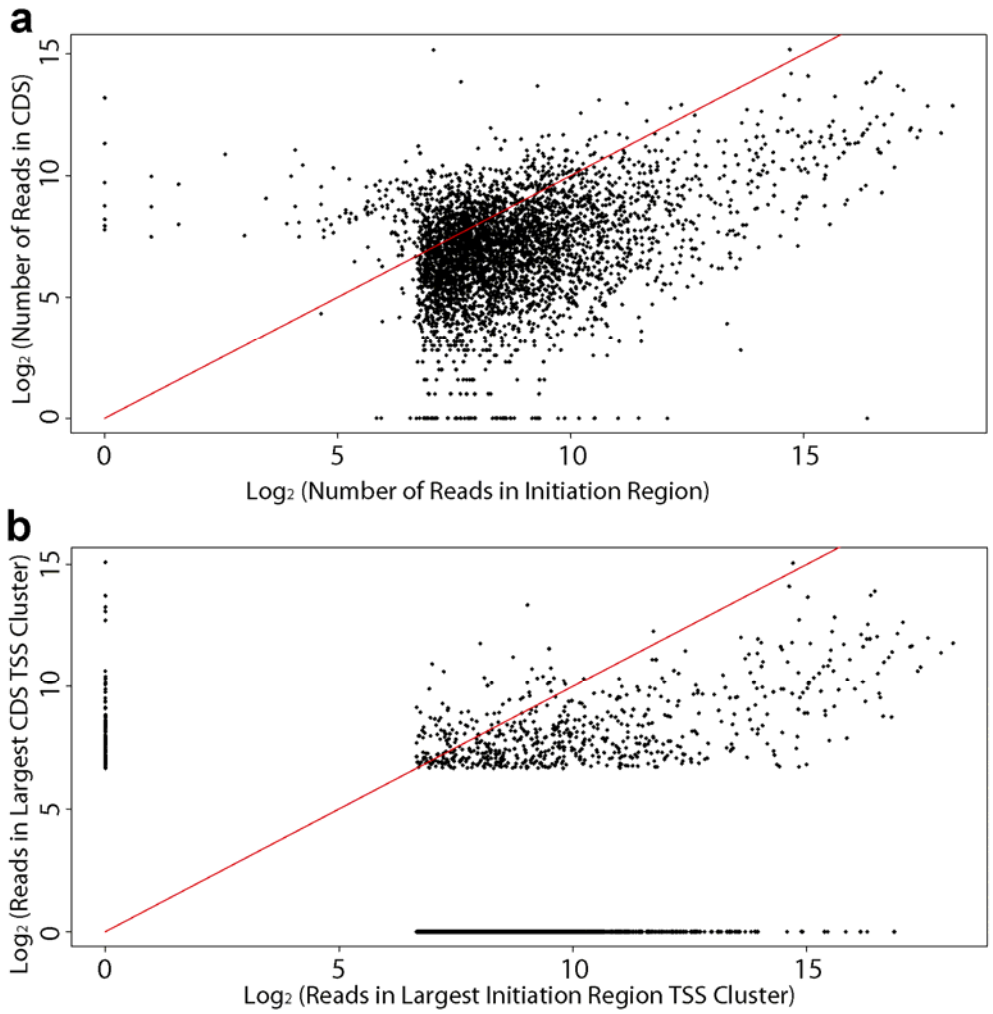


Figure 65: Comparison of the 5′ TSS reads mapped to the CDS and initiation regions

TSS reads or clusters mapped in close proximity to annotated genes were broadly divided into two categories. A TSS read (or cluster) was classified as "CDS" if it mapped to a coding region. Likewise, the TSS reads were defined as "initiation region" if they overlapped with a TSS, fall within ≤250nt upstream of a TSS or within the 5′ UTR. Direct comparison between the two TSS categories was made by using either **(a)** the total number of 5′ reads or **(b)** the largest cluster of each category. To reduce potential complications due to stochastic noise, only genes with a TSS cluster containing more than 100 reads in either the "CDS" group or the "initiation region" group, or both, were considered in (b).

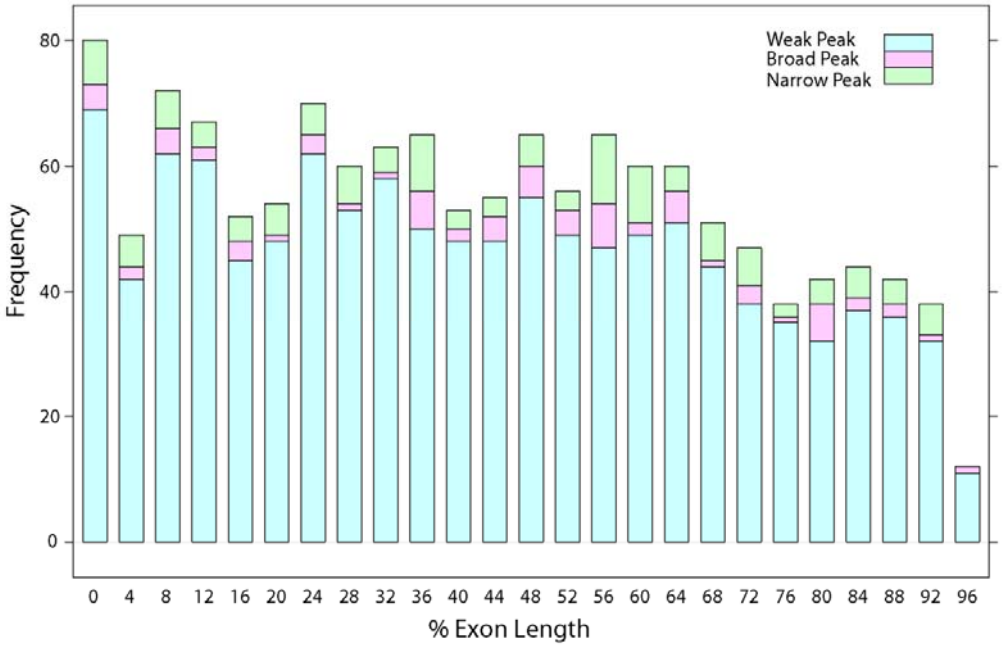


Figure 66: Distribution of internally capped TSS clusters across exons

The internal read clusters are defined as those containing more than 100 reads and are mapped to a coding region. The mode of each cluster was used to determine its relative location within an exon. The results were normalized by the exon length and further divided among three promoter classes (NP, BP and WP). To accommodate reads spanning across the exon junction, we first aligned the paired-reads to the genome, and then the transcriptome when the 5′ read could be mapped to the genome and the 3′ read was not. Details are described in the Methods section.

6.4 Discussion

6.4.1 The PEAT Method Leads to More Reliable 5' Reads

We have demonstrated that PEAT has several advantages over conventional technologies for the high-throughput identification of TSSs. The paired-reads allow for more accurate mapping of the raw data, and help compensate for alignment mistakes

caused by sequencing errors. This is especially evident when comparing the mapping efficiency of our data to those recently published for *D. melanogaster* (Ahsan et al. 2009) (Table 10). In addition, the 3′ read provides additional information for the local transcript structure. Thus, it can help link the distant TSSs to annotated genes and resolve internally capped transcripts at specific processing stages.

6.4.2 Distinct Promoter Types Exist in *Drosophila*

High-quality PEAT data allowed us to identify 3 distinct types of transcription initiation patterns. Extending previous observations based on *Drosophila* ESTs in Chapter 3 (Rach et al. 2009), the well studied location specific motifs (TATA, INR, DPE, MTE) were observed in NP promoters. In contrast, the WP promoters, which lack bias towards a specific location for initiation, were found enriched for the DRE motif. The presence of WP promoters in *Drosophila* is an interesting observation by itself as broad promoters in mammals are enriched for CpG islands (Carninci et al. 2006), a genomic feature not present in the fly. Notably, CpG islands and DRE are associated with housekeeping genes in human (Carninci et al. 2006) and fly (Engstrom et al. 2007) respectively, indicting functional conservation of WP promoters in diverse organisms. Moreover, ChIP data support the notion that distinct complexes may be associated with WP and NP promoters in fly. As their initiation patterns suggest, BP promoters contain a combination of both the motifs seen in the other two classes. However, it is unclear

whether this is a consequence of different complexes recognizing the same regulatory region, or if this occurs at different transcripts under the same condition. Our mixed-stage embryonic sample contains both maternal and zygotic transcripts, and vertebrate transcription in oocytes has recently been shown to depend on stage-specific basal transcription initiation complexes (Akhtar and Veenstra 2009; Gazdag et al. 2009).

6.4.3 Read Clusters Found Within Coding Regions Have Their Own Unique Features

As it has been reported in mammals (Carninci et al. 2006), we found a large number of 5′ capped transcripts that at first glance appear to initiate from the coding region of a gene. We provide multiple lines of evidence that these internally capped transcripts are largely not the result of alternative transcription initiation. Instead, they are likely derived from post-transcriptional processing events, as suggested by a previous mammalian study (Fejes-Toth et al. 2009). While the starting location of these transcripts was not supported by the presence of any core promoter motifs, we detected a distinct 'TC' di-nucleotide motif frequently occurring at the 5′ end, suggesting the possibility that a distinct set of factors may recognize these locations for the "recapping" of these transcripts. Consistent with the earlier study (Fejes-Toth et al. 2009), we found that recapping sites are uniformly distributed across the internal exon except at its extreme 3′ end. This coincides with the exon junction complex (EJC), which is deposited 20-24nt upstream of splicing junctions. Since both the early report and our study suggest

that internally capped transcripts are likely to be derived from processed (or spliced) transcripts, we speculate that the depletion of the recapping site at the end of the exon may reflect the competition between the EJC and recapping machinery. If this is the case, it would suggest that recapping takes place either in the nucleus (where EJCs are deposited) or during the first round of translation (where EJCs are stripped off) (Reichert et al. 2002; Shibuya et al. 2004). Further investigations are required to elucidate the biogenesis and functional significance of this novel class of transcripts.

Lastly, this study is focused on initiation sites of long polyadenylated transcripts. This explains why we did not observe promoter-associated non-coding transcripts, which have been reported in other species (Core, Waterfall, and Lis 2008; Kapranov et al. 2007; Seila et al. 2008), or the short transcripts associated with polymerase stalling (Muse et al. 2007; Zeitlinger et al. 2007). Our focused experimental design allowed us to more reliably determine the correlation between initiation patterns and core promoter sequence motifs. For instance, an earlier study using total RNA detected a large number of transcribed fragments (transfrags) that are well upstream of known TSSs and correlate in expression with the downstream genes (Manak et al. 2006). We showed that such distant TSSs are relatively rare for polyadenylated and capped transcripts, and are unlikely the initiation sites for known downstream transcripts. Although we cannot rule out that the observed differences are due to stage variation (mixed stage library vs. several 2hr windows), it is suggestive that these transfrags are

not polyadenylated or capped, or both; and that they may represent instances of a class of regulatory RNAs (e.g. promoter associated long RNAs, PALRs) in the fly transcriptome. Further efforts are required to profile and characterize different classes of RNA to dissect the complexity and plasticity of eukaryotic transcriptomes.

7. Nucleosome Organization and Chromatin Structure Reflect Divergent Strategies for Transcription

Elizabeth Rach conceived of all of the work in this chapter. She performed the *Drosophila* analysis, while the comparative human results were contributed by Deborah Winter from Dr. Terry Furey's lab. The work was submitted to *Nature Structural and Molecular Biology* in March 2010.

7.1 Introduction

Nucleosomes are the critical factors in chromatin formation and organization and thus play a role in regulating the accessibility of the DNA to transcription factors (TFs). As a result, transcription start sites (TSSs) are often located in the vicinity of nucleosome free regions (NFR), followed by a periodic pattern of nucleosomes downstream (Mavrich et al. 2008; Schones et al. 2008) . Nucleosomes containing H2 and H3 histone variants have been shown to provide particularly strong signals for the beginnings of genes in eukaryotes (Jin et al. 2009; Mavrich et al. 2008; Raisner et al. 2005), as they are preferentially incorporated in or near areas of active transcription. Data on frequent modifications to the N-terminal histone tails have provided support for a histone code specifying functional domains in the genome; for instance, the tri-methylation of H3K4 has been shown to mark the promoter regions surrounding TSSs (Barski et al. 2007). In addition, individual instances of insulator elements have been

shown or suggested to play a role in chromatin remodeling near promoter regions (Fu et al. 2008; Tsukiyama, Becker, and Wu 1994) .

Studies which explore nucleosome organization and histone modifications have largely treated all promoters as one group, and are typically based on a high-level view of promoters as provided in annotation databases. The development of high-throughput sequencing strategies, which generate libraries of millions of 5′ complete sequence tags from capped mRNAs, have provided a more fine-grained picture of transcription initiation. In particular, the application of Cap Analysis of Gene Expression (CAGE) technology (Carninci et al. 2006) has led to comprehensive sets of mammalian 5′ complete tags. Each of the tags corresponds to an initiation event, and mapping the tags to the genome has identified distinct initiation patterns characterized by broad and peaked tag clusters. We have recently extended this methodology to the Paired End Analysis of TSS (PEAT) protocol (Ni et al. 2010), mapping millions of paired reads in *Drosophila melanogaster* embryos. Analysis of these data showed that both fruit fly and mammalian promoters had comparable initiation patterns, with distinct associations to promoter motifs and functional roles.

Given that these distinct initiation patterns are widely conserved throughout eukaryotes, it may be surprising that no study as of yet has examined whether the narrow or wide distribution of initiation events may be caused by, or correlated with, the accessibility of DNA defined by the chromatin structure. In this work, we show that

distinct promoter classes, solely defined by their initiation patterns, have markedly different associations with nucleosome organization and histone modifications. These observations are further supported by distinct associations to recently defined *Drosophila* insulator classes (Negre et al. 2010). Our findings are conserved between humans and flies and thus strongly suggestive of two basic divergent strategies for gene regulation in eukaryotes.

7.2 Materials and Methods

7.2.1 Selection of Fruit Fly Transcription Start Sites

We used a dataset of 1,260 Narrow Peak (NP), 753 Broad with Peak (BP), and 2,041 Weak Peak (WP) promoters from *D. melanogaster*, determined by clustering of >10 mio. aligned 5′ capped paired-end sequence tags from 0-24 hour mixed stage embryos (Ni et al. 2010). The promoters were supported by at least 100 tags from *D. melanogaster* 0-24 hour mixed staged embryos previously mapped in Release 5.14 using the PEAT technology (Ni et al. 2010). This methodology starts from polyadenylated transcripts and generates pairs of a 5′-capped and an internal 3′ reads which allow for amore reliably mapping of transcription initiation sites to known transcripts. Promoters are classified by means of two features, genomic span of initiation events (as defined by the size of distinct 5′ tag clusters), and localization of initiation. For NP promoters, tag clusters have to be smaller than 25nt, and at least 50% of tags align at the peak location (defined

as the mode of the cluster +/- 2nt). BP promoters exceed the 50% tag cutoff at the mode, but are spread out over a genomic range > 25nt. WP promoters are those which meet neither genomic span nor peak location cutoffs; they do however still show a distinct albeit lower peak, frequently associated with the presence of a minimal initiator sequence motif.

Table 17: The Conversion Statistics for Mapping the Affymetrix Tiling Arrays From Release 4 to Release 5

Column 1 notes the chromosome, and columns 2 and 3 list the number of tiles in Release 4 and Release 5, respectively. Column 4 contains the number of tiles that were removed because they were mapped to multiple locations, or did not map to within 5 bp of the Release 4 tile size. Column 5 and column 6 cite the genomic locations of the first and last tiles in Release 5. Promoters identified using PEAT that were located outside of the scope of the Release 5 Affymetrix tiling array were excluded from the evaluation of temporal utilization using the 2-hr time course.

Chromosome	No. Rel 4	No. Rel 5	No. Removed	First (bp)	Last (bp)
2L	587,831	587,814	17	88	22,415,387
2R	534,929	534,838	91	380,463	21,146,537
3L	617,005	616,891	114	18,781	23,817,683
3R	740,120	740,120	0	15	27,904,884
4	28,284	28,284	0	578	1,281,978
X	558,566	558,529	37	18,910	22,422,262
Total	3,066,735	3,066,476	259		

The temporal activity of each promoter was determined through Affymetrix tiling array data that measured RNA levels every 2 hours during the first 24 hours of *D. melanogaster* embryogenesis (Manak et al. 2006). Published tile locations referred to release 4 and were converted to release 5 (see Table 17) using the Flybase Coordinate Converter (Wilson, Goodman, and Stretlets 2008). Due to differences in sequence, the 35bp resolution of the 25bp tiles was not fully conserved across releases for the ends of the chromosomes (see Table 17). As a result, promoters mapping to heterochromatin were only evaluated if they were in the scope of the array. Tiles that did not map uniquely or mapped to regions that differed more than 5bp in size from release 4 were excluded from the analysis. The utilization of promoters at each time point was evaluated as described in Chapter 4 (Rach et al. 2009), with the difference that all promoters in the set were evaluated regardless of distance between TSSs. Median fluorescence of three tiles downstream of a reference point was subtracted from three tiles upstream of the reference point, and promoters were deemed utilized if the differences were above previously determined time point specific thresholds (see Table 18) (Rach et al. 2009). Based on the narrow tag distribution within NP promoters, the mode location was used as the reference point for this class, and the start of the read clusters determined by Ni et al (Ni et al. 2010) were used for BP and WP promoters.

Table 18: False Positive Rates of Expressed Transcript Calls at TSSs

For each time point (column 1) corresponding to a 2 hour interval (column 2), a previously determined difference threshold (column 3) was used to determine false positive rates (column 4) for TSS utilization from background noise as in Chapter 4 (Rach et al. 2009). FP rates were consistently below 0.04.

Time Point	Hours	Difference Threshold	False Positive Rate
1	0-2	36.5	0.02
2	2-4	23.5	0.01
3	4-6	21	0.03
4	6-8	23.5	0.02
5	8-10	29.5	0.02
6	10-12	35	0.02
7	12-14	24.5	0.02
8	14-16	32	0.03
9	16-18	28	0.02
10	18-20	21	0.03
11	20-22	23	0.01
12	22-24	18.5	0.02

The significance of active promoter calls was evaluated by repeating the analysis on three sets of 1,000 randomly selected intergenic sites. Sites with values above a given threshold were counted as false positives (column 4). Relative false positive rates were averaged across the three sets (see Table 18). We obtained false positives rates ranging from .01 to .03 across all 2-hr time points (see Table 18); these rates were more stringent than previously observed and in Chapter 4 of .02-.046 (Manak et al. 2006; Rach et al. 2009). The Affymetrix tiling arrays measure expression throughout the genome and without polyA+ selection. Thus, the true false positive rates may be even smaller than .01 to .03, as the random intergenic sites that were chosen could reflect true transcription of e.g. unannotated short RNAs.

7.2.2 Scoring Fruit Fly Nucleosome and Regulatory Factor Profiles

Mavrich et al determined nucleosome positions by deep sequencing of MNase digested DNA associated with nucleosomes containing the H2A.Z histone variant, as well as by tiling array hybridization of bulk and pol-II associated nucleosomes. The published data had been processed to retain peaks above background, reflecting the midpoints of nucleosomes. From this data, we calculated normalized nucleosome occurrences for the H2A.Z, bulk, and pol-II bound data by first determining distances of the TSSs from the nucleosome midpoints with respect to the orientation of transcription, and adding them into 10bp non-overlapping bins. The moving average of five neighboring bins within the window from -1kb to +1kb was then normalized to the number of nucleosome occurrences per 500 TSSs. Enrichments are contrasted with averaged results of profiles on three sets of 1,000 random intergenic (RI) sites.

H3K4 methyl marks and insulator binding profiles were measured by hybridization to tiling arrays that were acquired from the modENCODE repository, and cumulated into 100bp bins relative to TSS locations. The moving average over three neighboring bins within -1kb to +1kb was normalized to the number of occurrences per 500 TSSs. The same strategy was again repeated on sets of random intergenic sites.

For a complete summary of data sources, see Table 19.

Table 19: Summary of Data Sources Used for Promoter Comparisons in Fly

The data type (column 1), and publication source (column 2) are listed with the total number of Release 5 locations (column 3). The sample source (time window during embryogenesis) and the type of experiment are summarized in columns 4 and 5, respectively.

D.mel Data	Publication	No.	Embryo (h)	Exp Type	Compared TSSs (h)
TSS	Ni 2010	4,054 promoters	0-24	PEAT	0-24
Bulk	Mavrich 2008	415,119	0-12	ChIP-chip	0-12
H2A.Z	Mavrich 2008	112,750	0-12	ChIP-seq	0-12
GAGA	Mavrich 2008	44,684	All	binding sites	0-12
polII Mpeaks	Mavrich 2008	2,832	Stage 14	ChIP-chip	12-16
polII nuc	Mavrich 2008	82,969	0-12	ChIP-chip	0-12
CTCF Nterm	ModENCODE Aug-8-08, DCCid 770	3,833	0-12	ChIP-chip	0-12
CTCF Cterm	ModENCODE Aug-8-08, DCCid 769	4,432	0-12	ChIP-chip	0-12
GAF	ModENCODE Aug-8-08, DCCid 23	6,438	0-12	ChIP-chip	0-12
Mod(mdg4)	ModENCODE Aug-8-08, DCCid 24	3,975	0-12	ChIP-chip	0-12
Su(Hw)	ModENCODE Aug-8-08, DCCid 27	4,779	0-12	ChIP-chip	0-12
BEAF-32	Négre 2010, GEO: GSM409067	4,710	0-12	ChIP-chip	0-12
CP190	Négre 2010, GEO: GSM409068	6,653	0-12	ChIP-chip	0-12
polII	Négre 2010, GEO: GSM409077	7,214	0-12	ChIP-chip	0-12
H3K4me3	Négre 2010, GEO: GSM409075	7,043	0-12	ChIP-chip	0-12
Expression	Manak 2006	3,066,476 probes	0-24	Tiling Array	0-24

7.2.3 TSS Cluster Identification From Human CAGE Tags

For a comparable analysis in *Homo sapiens*, we started from the published alignments of 29 million tags generated by the FANTOM consortium. CAGE tags were grouped into clusters using the same strategy and parameters as published previously for fruit fly (Ni et al. 2010). Clusters were assigned to genomic location, and we here analyzed those within initiation regions, which included annotated 5'UTRs and 250bp upstream of the annotated TSS in the UCSC genome browser (27% of all clusters). Promoters of clusters in the initiation region were further classified as NP (1205), BP (1588), and WP (7160) based on the shape of their tag distributions (Ni et al. 2010). The modes of the tag distributions were used as representative TSS locations for all promoter classes.

7.2.4 Scoring Human Nucleosome and Regulatory Factor Profiles

The nucleosome occupancy score for H2A.Z, H3K4 methylation, and bulk profiles was calculated according to Schones et al, using raw short aligned reads mapping to 5' or 3' nucleosome boundaries (Schones et al. 2008). We divided each somatic chromosome into 10bp non-overlapping windows, and read counts for a window were calculated by summing the number of reads that aligned in the 80bp upstream (on the sense strand) or 80bp downstream (on the anti-sense strand) windows, assuming that 5' and 3' reads mapping to the ends of the same nucleosome would be

~140-160 nt apart. Promoters were analyzed in windows from -1kb to +1kb of the TSSs identified by tag clustering, and to reduce the noise in the bulk data, promoters with outlier read counts less than 3 or greater than 3,000 were removed from the analysis. A raw nucleosome occupancy score was determined for each promoter window by averaging the read counts across all of the individual promoters within one pattern (NP, BP, and WP). A moving average over five windows of raw nucleosome occupancy scores was taken for each promoter pattern to produce the smoothed nucleosome profiles shown. As in fly, window scores thus reflected nucleosome midpoints; unlike in fly, profiles are based on the complete read data instead of only midpoints as determined by local maxima. A set of 5,000 random intergenic sites was chosen across Chromosome 1 for which nucleosome profiles were determined akin to that of the promoters.

For pol-II, DHS, and CTCF profiles, raw read data was assigned to 10bp non-overlapping windows regardless of strand. Within each promoter pattern, the read counts were averaged for windows covering +/- 1kb with respect to their locations from the TSS, and a moving average over five windows was used for smoothing, resulting in the average read density shown in the figures. The same steps were applied to the set of random intergenic sites from Chromosome 1.

For a complete summary of data sources, see Table 20.

Table 20: Summary of Data Sources Used for Promoter Comparisons in Human

The data type (column 1), and publication source (column 2) are listed with the total size of the dataset (column 3) and the cell type in which it was generated (column 4). Column 5 refers to the type of experiment performed.

H.sap Data	Publication	No.	Cell Type	Exp Type
TSS	Kawaji 2009 (FANTOM4)	29 mil tags	127+ cell types	CAGE
Bulk	Schones 2008	9,097,773	CD4+ T-cells	MNase-seq
H2A.Z	Barski 2007	780,887	CD4+ T-cells	MNase/ChIP-seq
DHS	Boyle 2008	776,108	CD4+ T-cells	DNase-seq
CTCF	Barski 2007	139,700	CD4+ T-cells	ChIP-seq
polII	Barski 2007	478,748	CD4+ T-cells	ChIP-seq
H3K4me1	Barski 2007	498,034	CD4+ T-cells	MNase/ChIP-seq
H3K4me2	Barski 2007	432,927	CD4+ T-cells	MNase/ChIP-seq
H3K4me3	Barski 2007	4,472,422	CD4+ T-cells	MNase/ChIP-seq
Expression	Boyle 2008	14.6 mil probes	CD4+ T-cells	Tiling Array

7.2.5 Stratification by Human Expression Levels

The log values of gene expression from NimbleGen tiling arrays for $CD4^+$ T-cells generated in an earlier study (Boyle et al. High-resolution mapping and characterization of open chromatin across the genome 2008) were mapped to corresponding TSSs via associated genes (see Figure 67). The log(expression) values of all genes, regardless of promoter pattern, were plotted and divided into four groups. As in a previous study, we declared genes below a cutoff of 4.5 as "silent", and divided the remaining genes evenly into three groups. Consequently, there were 837 genes with values below 4.5 that had 'no' expression, 3,039 genes above 4.5 and below 6.662 that had 'low' expression, 3,038

genes above 6.662 and below 8.258 that had 'medium' expression, and 3,039 genes with values higher than 8.258 that had 'high' expression. Within each expression group, the TSSs were then subdivided a second time according to their promoter pattern (NP, BP, WP). Expression levels across promoter patterns were thus based on the same cutoffs. Occupancy scores were then calculated as described above. As there were nearly four times more promoters associated with genes having high/medium/low expression than those with no expression, occupancy profiles for 'no' expression are less smooth.

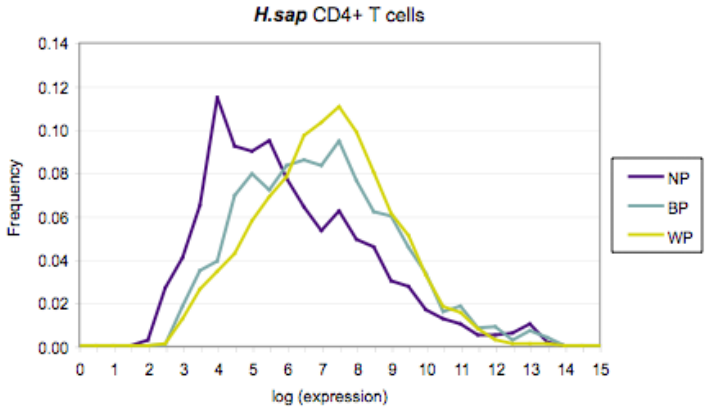


Figure 67: Expression Levels of Human Genes by Promoter Class

Gene expression intensities from NimbleGen tiles (Boyle et al. High-resolution mapping and characterization of open chromatin across the genome 2008) were assigned to human TSS clusters (see Methods), and their log(expression) values were binned separately for each promoter class and normalized to relative frequencies. BP and WP promoters had nearly identical expression, while NP promoters showed a skew towards lower expression.

7.3 Results

7.3.1 Fruit Fly Promoters Have Equal Expression During Embryogenesis

Based on previous studies in mammal (Carninci et al. 2006) , our recent work in *Drosophila* embryos has led to the definition of three main transcription initiation patterns defined by the size of the initiation region and the distribution of initiation events within each region (Ni et al. 2010). Narrow Peak (NP) promoters have high occurrences of initiation at one location, contain canonical position-specific core promoter motifs such as the TATA box, and are utilized for developmental regulation and tissue-specific functions. Conversely, initiation in Weak Peak (WP) promoters is more broadly distributed over a larger genomic span. WP promoters are associated with distinct sequence elements but lack the canonical core promoter motifs in fly, largely coincide with CpG islands in mammals, and are found in housekeeping genes (Engstrom et al. 2007; Rach et al. 2009). Broad with Peak (BP) promoters display a combination of features from both the NP and WP promoters.

D. melanogaster promoters were defined based on mixed stage embryonic libraries, and we used tag clusters with at least 100 tags. Since this is significantly above background, all promoters under investigation were therefore linked to transcripts present during embryogenesis (see Figure 68). The patterns of temporal utilization across all three promoter patterns showed a dip in utilization at time point 3 (hours 4-6) followed by an increase in utilization at time points 4-8, and a second smaller increase in

utilization at time points 9-12 (see Figure 68A). This pattern corresponds with previous work based on a careful analysis of EST-supported TSSs, as the dip at time point 3 corresponds to the degradation of maternally inherited transcripts and the production of new transcripts by zygotic transcription (Manak et al. 2006). When comparing across promoter classes, we saw higher numbers of BP and WP promoters utilized during the early to mid stages of development, and higher occurrences of the NP promoters utilized during the mid to late stages of development. This result also corresponded with previous associations of promoter classes with timing during development in Chapter 4 (Rach et al. 2009).

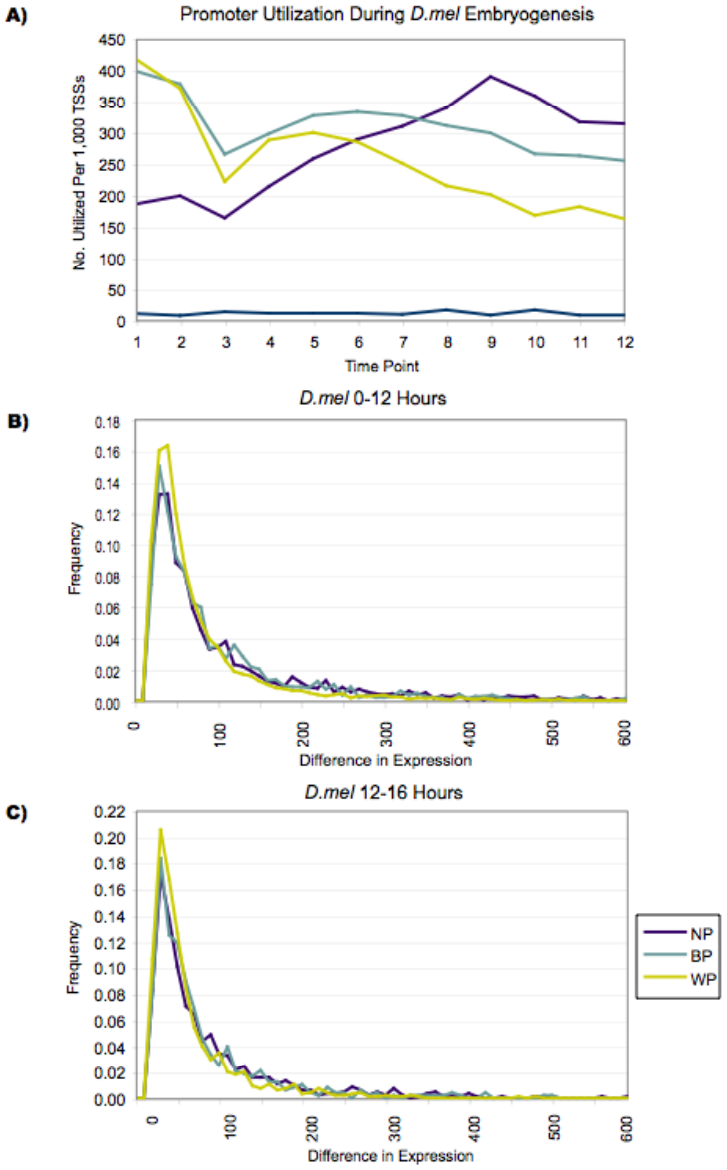


Figure 68: Fruit Fly Promoter Classes Show Different Temporal Trends at the Same Magnitude of Expression

The time points of utilization for each promoter were determined using the differences in median fluorescence intensity values of the Affymetrix tiling arrays as in Chapter 4 (Rach et al. 2009). The number of promoters with utilization at each time point were added by pattern and normalized per 1,000

TSSs. **(A)** The overall progression of expression agreed with previous results: higher numbers of BP and WP promoters were utilized during the earlier stages of embryogenesis, while the opposite was true for NP promoters. **(B)** Promoters with utilization in at least one time point from 0-12 hours were assigned to expression levels based on array fluorescence (differences in median fluorescence of tiles downstream of a TSS vs. upstream, discretized in bins of size 10). Promoter numbers in each bin were divided by the total number of differences, resulting in the frequency of expression as shown. A line graph was used to smoothly join the discrete bin densities. While quantities of promoter patterns changed throughout embryogenesis (A), the distribution of expression levels was the same across all promoters. **(C)** The expression analysis was repeated for promoters with utilization in at least 1 time point from 7 to 8 (hours 12-16, to match pol-II occupancy data from developmental stage 12). Again, a similar distribution of expression levels from the tiling arrays is observed across all promoter patterns.

To restrict promoters to sets mapping to published nucleosome and binding data, we selected the set of 517 NP, 406 BP, and 1,054 WP promoters that were utilized during the first 12 hours of development. This set of promoters was thus supported by both the PEAT sequencing and Affymetrix tiling arrays, and their precise 5′ locations and exact timing of utilization were verified by independent experiments. We plotted the fluorescence differences at utilized promoters, and divided by the total number of promoters to obtain the expression distribution for each promoter class. Figure 68B shows that the expression is highly similar across all of the promoter classes. The WP promoters had slightly higher frequency of fluorescence differences from 30 to 60 and the NP and BP promoters had higher occurrences of the differences greater than 100. However, these observed differences did not significantly differentiate the expression

profiles of the promoter patterns. Thus, the quantities of NP, BP, and WP promoters that are utilized changed over the time course of fruit fly embryogenesis; however, for promoters with transcription above background, the level of expression was the same across promoter classes.

7.3.2 Promoter Classes Exhibit Distinct Nucleosome Organization

We first evaluated the positioning of nucleosomes containing H2A.Z with respect to the start sites, as this histone variant has been associated with clearer signals in promoters when compared to bulk nucleosomes (Mavrich et al. 2008). Fruit fly BP and WP promoters showed a significantly greater association with the organization of the H2A.Z nucleosomes than NP promoters (Figure 70A). BP and WP promoters also had a greater percentage of H2A.Z nucleosomes within 1kb of the TSS (Figure 69), and the spacing between H2A.Z peaks for BP and WP promoters was more consistent than for NP promoters. As NP promoters have the highest enrichment of TATA boxes, this explains the previous observation that TATA-containing promoters have a 'very fuzzy' H2A.Z nucleosome organization (Albert et al. 2007; Mavrich et al. 2008). Thus, not all promoters exhibit the same underlying nucleosome organization; rather, promoters with broader distribution of initiation are associated with a more clearly defined periodic nucleosome organization, whereas promoters with precisely positioned start sites are less organized.

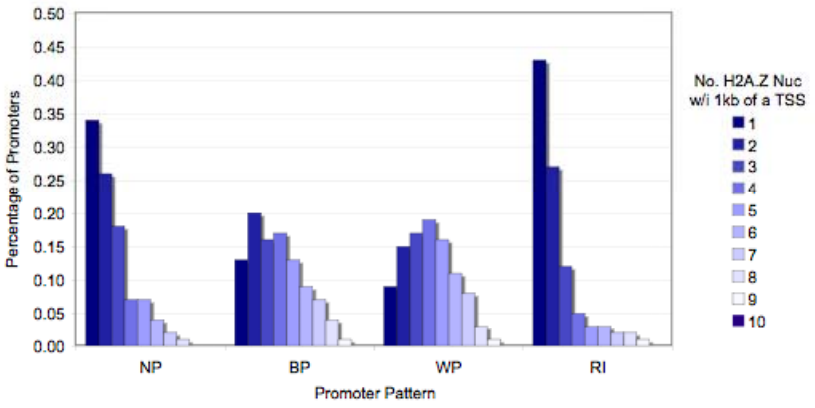


Figure 69: Density of H2A.Z Nucleosomes Is Higher in BP and WP Promoters Than in NP Promoters

The midpoints of all H2A.Z nucleosomes were taken from Mavrich et al and mapped to the locations of the 0-12 hour NP, BP, and WP promoters. There were 95% of WP and 89% of BP promoters that had at least one H2A.Z nucleosome within 1kb of a TSS, compared to 79% of NP and 71% of random intergenic sites. Greater differences in percentages were observed for BP and WP promoters with more than one nucleosome within 1kb of the TSS. This illustrates the stronger connection of BP and WP promoters to the positioning and quantity of H2A.Z nucleosomes within the immediate vicinity of the TSS.

When the locations of bulk nucleosomes in fruit fly were compared across promoter classes, an overall fuzzier signal was obtained. This is consistent with previous observations (Mavrich et al. 2008), and may partially be due to the resolution of the tiling arrays used to measure the fly bulk profiles. Still, a consistent difference between promoter patterns was observed (Figure 70B). For the NP promoters, the +1 nucleosome was observed at 125bp and the -1 nucleosome was observed at -180bp, slightly adjusting the previous numbers in fruit fly based on genome annotations rather than precise TSS

locations (Mavrich et al. 2008; Yuan et al. 2005). Despite differences in nucleosome occupancy, all fruit fly promoter classes exhibited the same levels of RNA polymerase II (pol-II) bound nucleosomes assayed in the same fruit fly embryos as the bulk nucleosomes (Mavrich et al. 2008), and the peak of pol-II in all three promoters occured at the location of the +1 nucleosomes (at +115bp; Figure 70C). An elevated level of pol-II bound nucleosomes further downstream was suggestive of active transcription.

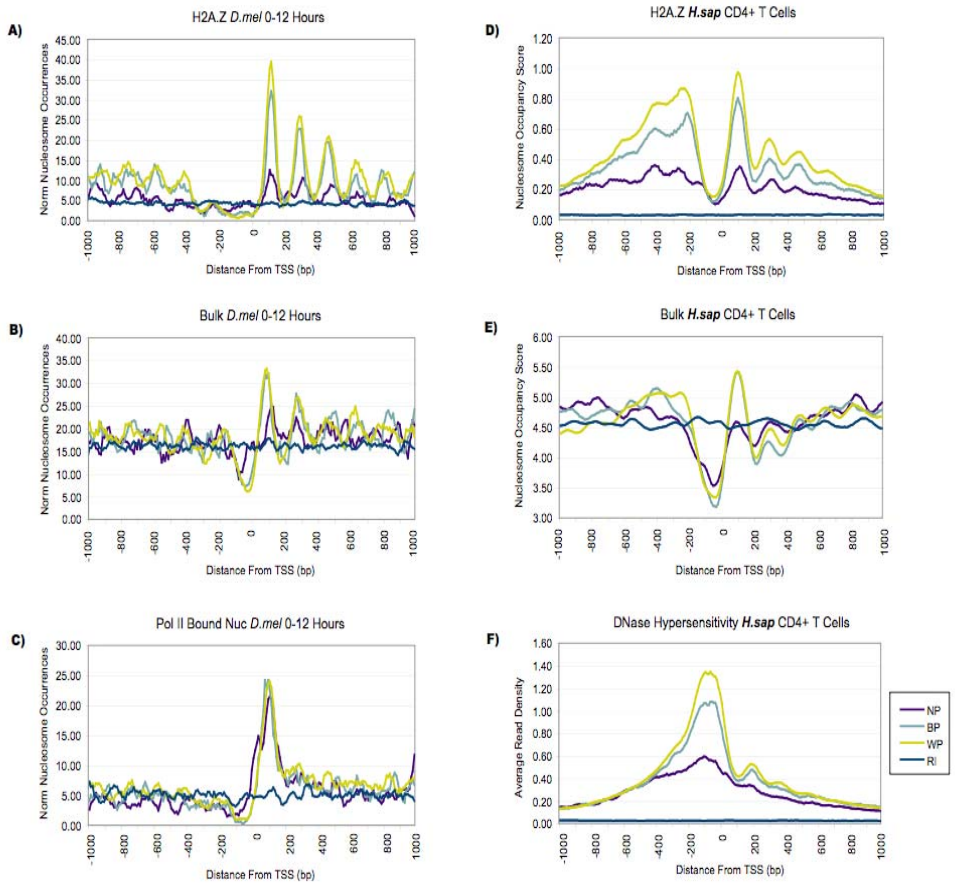


Figure 70: Nucleosome Organization is Promoter Class Specific

Profiles are based on promoters classified as Narrow Peak (NP), Broad with Peak (BP), and Weak Peak (WP), and show the region of –1 kb to +1 kb around the designated TSS. For fly analyses, promoters with active transcription in at least one time point from 0-12 hours of fruit fly embryogenesis were used (a set of 517 NP, 406 BP, and 1,054 WP promoters); for human, promoters were mapped and classified based on FANTOM 4 sequence tags (a set of 1,205 NP, 1,588 BP, and 7,160 WP promoters). As baseline, RI refers to average levels at random intergenic sites. **(A)** Fruit fly H2A.Z profiles show that BP and WP patterns had increased H2A.Z occupancy and organization. The +1 H2A.Z nucleosome occurred at 125bp, which is 10bp upstream of the previous estimate in fruit fly, and 65bp downstream of the +1 nucleosome in yeast (Mavrich et al. 2008; Yuan et al. 2005). The difference to previous estimate is likely a consequence of the more precisely mapped TSS data which underlie our profiles. Nucleosomes in BP and WP promoters had a more precise spacing, with an average separation of 170bp and deviations of up to 10bp, compared to a mean distance of 183 bp between H2A.Z peaks at NP promoters, with deviations of up to 33bp. **(B)** Differences between promoter classes are also apparent in *Drosophila* bulk nucleosomes profiles, with a slight shift compared to H2A.Z as observed in the original study (Mavrich et al. 2008) **(C)**. Despite differences in nucleosome association, the levels of pol-II bound nucleosomes are at comparable levels across promoter classes. Increased H2A.Z **(D)** and bulk **(E)** occupancy and spacing were also observed for human BP and WP promoters, akin to fruit fly. **(F)** DNase hypersensitive site profiles confirmed these associations, revealing a more accessible nucleosome-free region at BP and WP but not at NP promoters. The overall "sharper" appearance of the fruitfly profiles is likely the consequence of differences in generation and processing of nucleosome data as taken from the previous human and *Drosophila* studies (see Methods).

We next investigated whether these observations would be conserved across species. In particular, much of the available mammalian data on nucleosome organization and histone variants has been obtained from human CD4+ T-cells. This additionally allowed us to compare observations from the mixed cell population of the developing fly embryo with those from a single differentiated cell type. To maintain consistency across species, TSS clusters were determined from available human CAGE

tags in the FANTOM4 database (Kawaji et al. 2009) using the same methodology and parameters as in fruit fly (see Methods). We then compared the H2A.Z locations profiled in a previous study to the promoter patterns (Barski et al. 2007). Confirming the observations in fruit fly embryos, Figure 70D shows that BP and WP promoters had a higher association with H2A.Z nucleosome organization than NP promoters. The locations of the +1, +2, and +3 H2A.Z nucleosomes, and the 185 bp spacing between them, agreed with previous estimates (Fu et al. 2008; Tolstorukov et al. 2009). An apparent difference between the results for the two species was a lack of H2A.Z association at the -1 nucleosome in *Drosophila* as previously reported (Mavrich et al. 2008). However, this lack does not coincide with an overall lower level of bulk nucleosomes at this location (cf. Figure 70B). As this phenomenon was not observed in human, additional experiments would be beneficial to confirm this putative species-specific difference.

When plotting the bulk nucleosome locations in human (Schones et al. 2008), the +1, +2, and +3 nucleosome positions and the 185bp spacing between them were maintained (Figure 70E). The bulk NFR aligned to that of H2A.Z, and BP and WP showed a distinctly higher association with bulk nucleosome organization than the NP promoters, in particular at the +1 position. These observations are precisely mirrored by the prevalence of DNase hypersensitive sites (DHS) in human. Averaging over all promoters, previous studies reported that most promoters were accompanied by a DHS

site (Boyle et al. High-resolution mapping and characterization of open chromatin across the genome 2008). In agreement with the bulk nucleosome profiles, our results showed that only WP and BP promoters demonstrated a distinct peak, appearing twice as sensitive to DNase compared to NP promoters (Figure 70F). The maximum peak of sensitivity occurred ~100 bp upstream of the TSS, at the location of the NFR, with a second peak ~200bp downstream of the TSS, between the +1 and +2 nucleosomes. The DHS mirror image around the nucleosomes provides further evidence for the distinct nucleosome organization at BP and WP promoters. Despite the high divergence between fruit fly and human lineages, the conservation of promoter patterns and the associated differences with respect to nucleosome organization imply a functional importance of the underlying biological mechanism.

7.3.3 Promoter Classes Maintain Distinct Associations Across Expression Levels

Previous studies had consistently observed a correlation of nucleosome associations with the expression levels of genes (Boyle et al. High-resolution mapping and characterization of open chromatin across the genome 2008; Schones et al. 2008). To rule out the possibility that some observations could be explained by an overall lower activity of certain promoter classes (Figure 67), we divided the human CD4+ cell line data into four groups based on expression levels (high, medium, low, no), and compared each category to H2A.Z occupancy (see Methods). Figure 71A shows that BP and WP

promoters consistently had higher levels of H2A.Z occupancy than NP promoters regardless of expression levels. This demonstrates that in spite of differences in expression levels, promoter classes have different relationships to nucleosome organization. In addition, while H2A.Z enrichments have been reported to be present in promoters of both active and inactive genes in yeast (Raisner et al. 2005), and while we also observe H2A.Z enrichments for BP and WP promoters regardless of expression level, the H2A.Z association disappears at NP promoters with low or no expression. Given that the class of NP promoters is smaller, it is possible that this phenomenon was simply averaged out in previous analyses (Barski et al. 2007; Raisner et al. 2005), which did not split the data by promoter class.

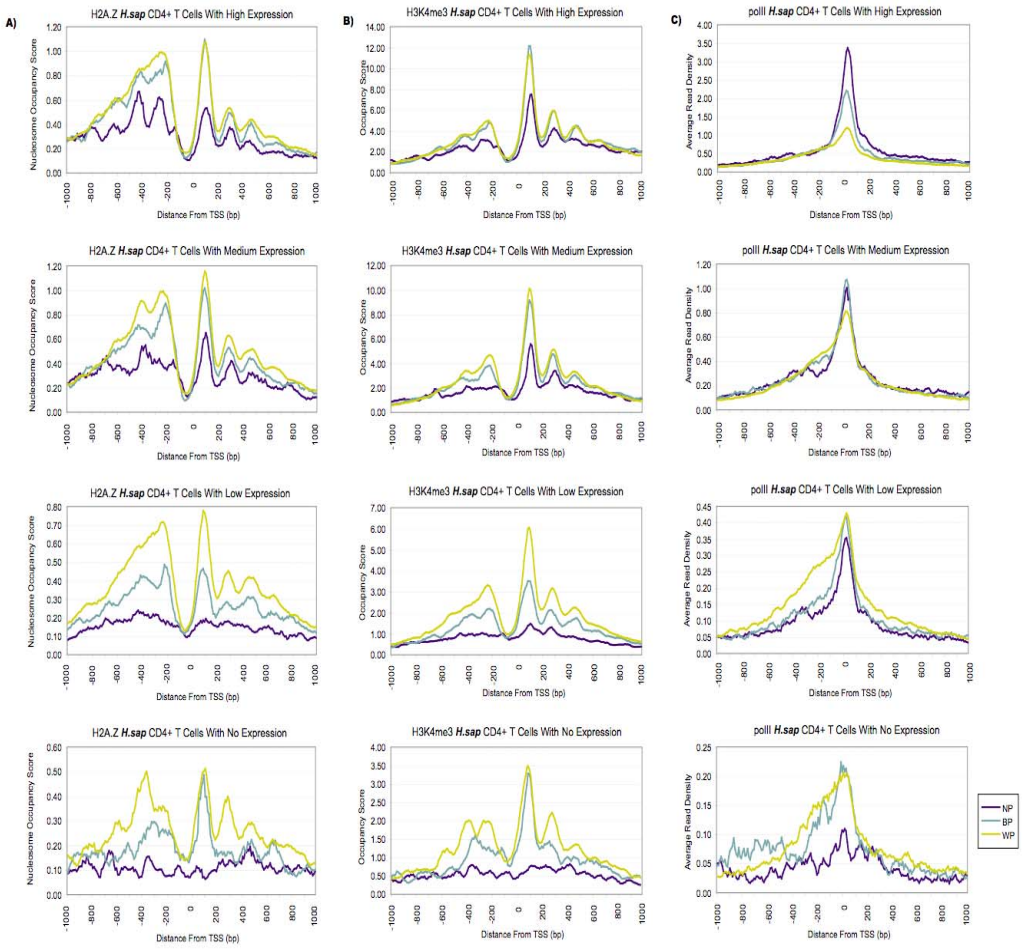


Figure 71: H2A.Z and H3K4 Trimethylation Profiles Separate Promoter Classes Even When Stratified by Expression Levels

Human promoters were separated into 4 classes based on expression levels of associated genes in $CD4^+$ T-cells. **(A, B)** Across all expression levels, BP and WP promoters showed greater enrichments in H2A.Z and H3K4 trimethylation than NPs. **(C)** Levels of pol-II binding showed the opposite trend, with NP promoters being much more occupied by pol-II at the TSS despite the much lower H2A.Z and H2K4me3 association. Note that the scale of nucleosome profiles varies by 2-3 fold across expression levels, and the scale of pol-II binding by more than an order of magnitude.

Studies have shown a coupling of H2A.Z with H3K4 methyl marks at TSSs (Barski et al. 2007; Wang et al. 2008). To validate this association across promoter classes and expression levels, we matched NP, BP, and WP promoters with human H3K4 methylation data (Barski et al. 2007). Figure 71B shows that the positioning of dips and peaks for the H3K4me3 signals across all promoters and expression levels corresponded with the positioning and levels of H2A.Z nucleosomes. Notably, BP and WP promoters had consistently higher H3K4me3 occupancy than NP promoters, and this also held for mono-and di-methylation (Figure 72). In contrast to this observation, NP promoters had a strikingly higher density of pol-II at high expression levels than BP and WP promoters (Figure 71C). Therefore, the lower levels of H2A.Z and H3K4me3 did in fact correspond to an increased presence of the polymerase. These higher levels of polymerase occupancy at NP TSSs, stratified by expression level, suggest a possible role for pol-II stalling frequently observed in more specialized cell types (Adelman et al. 2009; Nechaev et al. 2009).

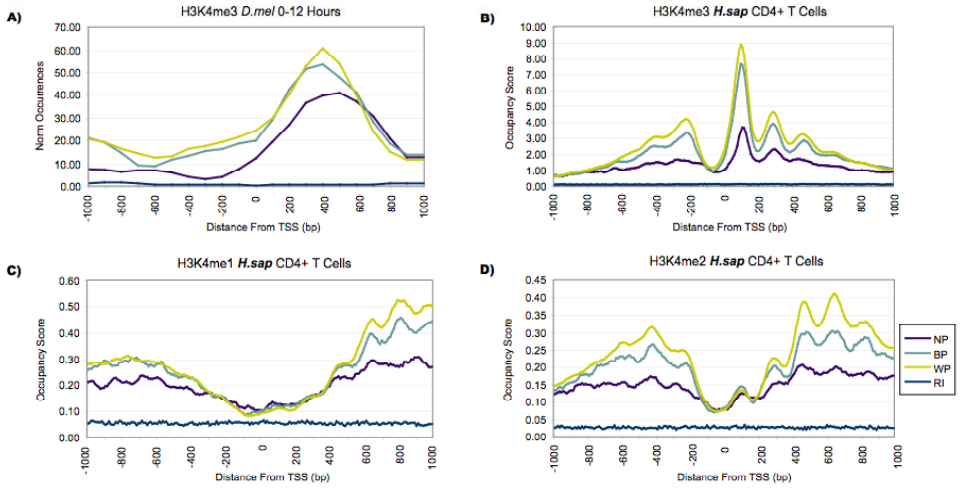


Figure 72: WP and BP Promoters Have Stronger Association to H3K4 Methylation

(A, B) Average profiles of H3K4me3 occupancy in *Drosophila* and human promoters show an overall similar pattern. **(C, D)** The lower association of H3K4 methylation for NP promoters is retained in human H3K4me1 and H4K4me2 profiles, which consistently show relative enrichments further within transcribed regions, but only for WP and BP promoters.

For *Drosophila*, the data are obtained from mixed embryo population, which makes it hard to investigate the influence of expression in detail. However, expression levels observed for transcripts associated with the different promoter classes are highly similar (cf. Figure 68), and an analysis of ChIP-chip data (Negre et al. 2010) on H3K4 tri-methylation shows a pattern consistent with human. Furthermore, while pol-II binding profiles in 0-12 hr early-stage embryos are fairly similar across classes, the profiles at a

later stage (stage 14, when cells become more differentiated) mirror the human results, with higher pol-II association to NP promoters (Figure 73).

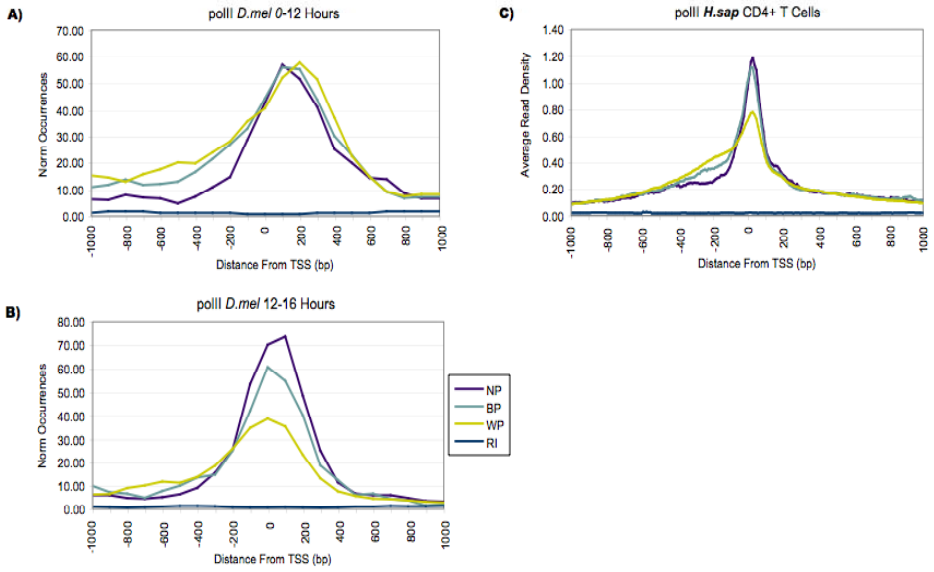


Figure 73: *Drosophila* NP Promoters Show Higher Levels of Pol-II Binding at Later Stages of Development

(A) Binding levels of pol-II obtained from 0-12 hr mixed-stage embryos showed comparable levels across promoter classes and agreed with pol-II associated nucleosomes (Figure 70C). **(B)** Later in development, NP promoters had noticeably higher levels of pol-II than BP and WP promoters, differing from the 0-12 hr profile but in agreement with the average human profile **(C)** obtained from differentiated CD4+ cells. 520 NP, 287 BP, and 587 WP *Drosophila* TSSs with utilization in at least one time point from 7 or 8 (hours 12-16) were retained from the full set, to map to the pol-II MPeak binding locations (embryonic stage 12) generated by Pugh et al (Mavrich et al. 2008).

7.3.4 Insulator Classes Coincide With Initiation Patterns

Insulators demarcate differentially expressed genes, disrupt the communication between enhancers and promoters, and prevent the spreading of chromatin domains. Individual instances of insulator elements have been shown or suggested to play a role in chromatin remodeling near promoter regions (Fu et al. 2008; Tsukiyama, Becker, and Wu 1994). We therefore investigated if insulator elements, just as nucleosome organization and chromatin state, may support the existence of different basic promoter classes.

The CCCTC-binding factor (CTCF) is one of the most prominent insulator proteins that is widely conserved across species (Smith et al. 2009). It is known to interact with pol-II, and has been implicated in the assistance of nucleosome positioning around its binding sites in human (Chernukhin et al. 2007; Fu et al. 2008). In particular, it is enriched at locations of H2A.Z and H3K4 methylation in human (Barski et al. 2007). Supporting this, CTCF showed a higher association with BP and WP promoters than NP promoters (Figure 74A) (Barski et al. 2007). The CTCF profile reached a maximum at -125bp upstream of the TSS. This organization places CTCF in the proximity of the core promoter and just downstream of the -1 nucleosome. These results agree with previous work showing that nucleosomes enriched for the H2A.Z variant were well-positioned and flanked by CTCF (Fu et al. 2008). Concordant results were observed when *Drosophila*

CTCF (dCTCF) binding was evaluated (Figure 74B), albeit with a broader enrichment due to the lower resolution of the tiling array.

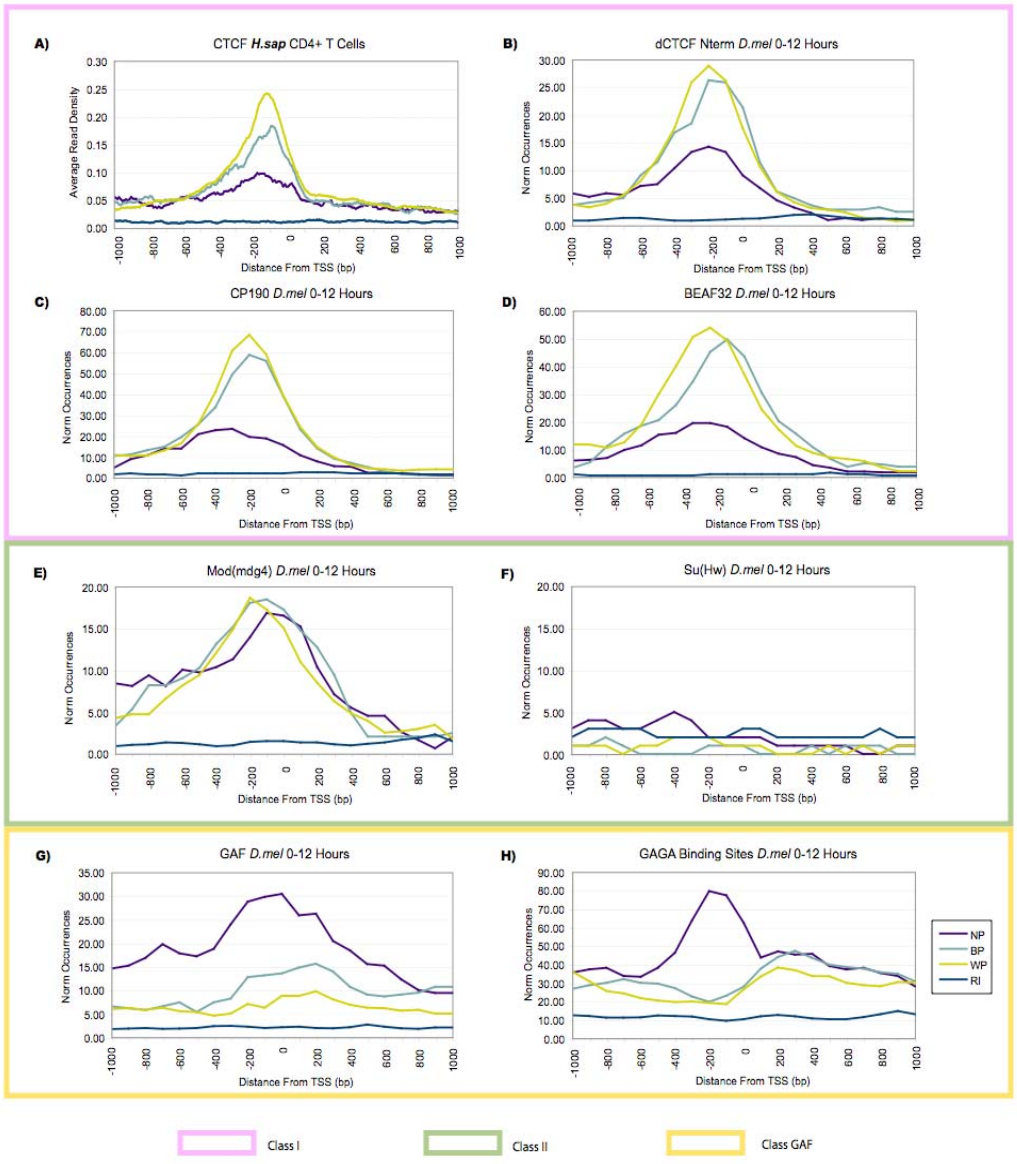


Figure 74: Insulator Classes Are Characteristic of Promoter Classes

Two classes of fruit fly insulators established in a previous study (Negre et al. 2010) were compared to promoters classes on embryonic data from 0-12 hr. **(A,C,D)** Class I insulators (dCTCF, CP190, and BEAF32) had higher occurrences as BP and WP promoters, which agreed with their frequent co-occurrence. **(B)** CTCF is conserved to human and supported the same pattern. (**E,F**) Class II insulators had equal occurrence across promoter classes, with Su(Hw) not being bound to proximal promoter regions as previously reported. **(G,H)** ChIP-chip profiles of the chromatin-remodelling transcription factor GAF, as well as presence of GAGA binding sites in the genome, showed a clear enrichment at NP promoters.

The availability of genome-wide data on insulator binding elements as part of the modENCODE project (Celniker et al. 2009) provided an opportunity to expand the observations made for dCTCF. The data was obtained from 0-12 hr mixed stage embryos, i.e. from the same material as the nucleosome data analyzed above (Negre et al. 2010). Genomic analyses defined two classes of insulator elements in fruit fly based on co-occurrence of binding events, and showed significant associations with genomic properties such as proximity and organization of genes and cis-regulatory elements. dCTCF, CP190, and BEAF32 comprise the Class I insulator elements in fruit fly (Negre et al. 2010). In accordance with the frequent co-occurrence of their binding sites, the two other Class I insulators, CP190 and BEAF132, also showed specific enrichments in WP and BP promoters (Figure 74C,D).

Class II insulators in fruit fly are comprised of Su(Hw) associated proteins (Negre et al. 2010). Mod(mdg4) and CP190 have been shown to recruit Su(Hw) to the *gypsy* insulator, however, Su(Hw) is reportedly not enriched in promoters (Negre et al.

2010). Mod(mdg4) had an equal enrichment across all promoter classes, which suggests similar functional roles across promoters (Figure 74E). As expected, Su(Hw) was absent from all promoters (Figure 74F).

Lastly, we investigated the GAGA binding factor (GAF) which did not cluster with factors in either Class I or Class II insulators (Negre et al. 2010). GAF can regulate gene expression at multiple levels, including mediating promoter-enhancer interactions and insulating chromosomal position effects (Mahmoudi, Katsani, and Verrijzer 2002). For instance, at the *D. melanogaster hsp70* promoter, GAF works in combination with the Nucleosome Remodeling Factor (NURF) to disrupt histone octamers over the GAGA site (Tsukiyama, Becker, and Wu 1994) and promote pol-II pausing (Lis 1998). In the context of initiation patterns, we observed a prominent enrichment of GAF binding in NP promoters from -1400bp to +1100bp of the TSS (Figure 74G). When scanning promoters for matches to the GAGA sequence motif, NP promoters showed high levels of matches in a narrower area within the region bound by GAF, and BP and WP promoters had a pronouncedly lower level (Figure 74H) – i.e., the opposite of Class I insulators. Therefore, at least in the case of GAF, the preference for a particular promoter class does not necessarily reflect a dynamic state (such as expression level), but rather is statically encoded in the DNA sequence.

Taken together, proteins from the recently defined insulator classes and the GAGA binding factor clearly separated among the promoter classes. As the definition of

insulator classes was based on features of genomic organization, and the definition of promoter classes was based on initiation patterns agnostic to position and orientation of genes or cis-regulatory modules in their vicinity, this is a remarkable confluence and further confirms the notion of distinct fundamental promoter types.

7.4 Discussion

Many recent studies have reported on the chromatin structure in eukaryotic genomes, and identified stereotypical patterns of nucleosome organization and histone marks. While differences related to the expression levels of genes were observed, most of these studies assumed that all promoters fundamentally share the same chromatin organization, as reflected in the use of average profiles across all promoters. Our approach differs from current efforts (Lee et al. 2007) by allowing us to assess nucleosome positioning, periodicity and function from the basis of the transcription initiation machinery, and indicates that averaging over all promoters may have previously obstructed these distinctions. The high-throughput sequencing of 5' sequence tags has clearly shown that eukaryotic promoters separate into several classes defined by broad and narrow distributions of initiation events, and we have here established that promoters from different classes are characterized by distinct patterns of nucleosome organization, chromatin structure, and insulator preferences (Figure 75).

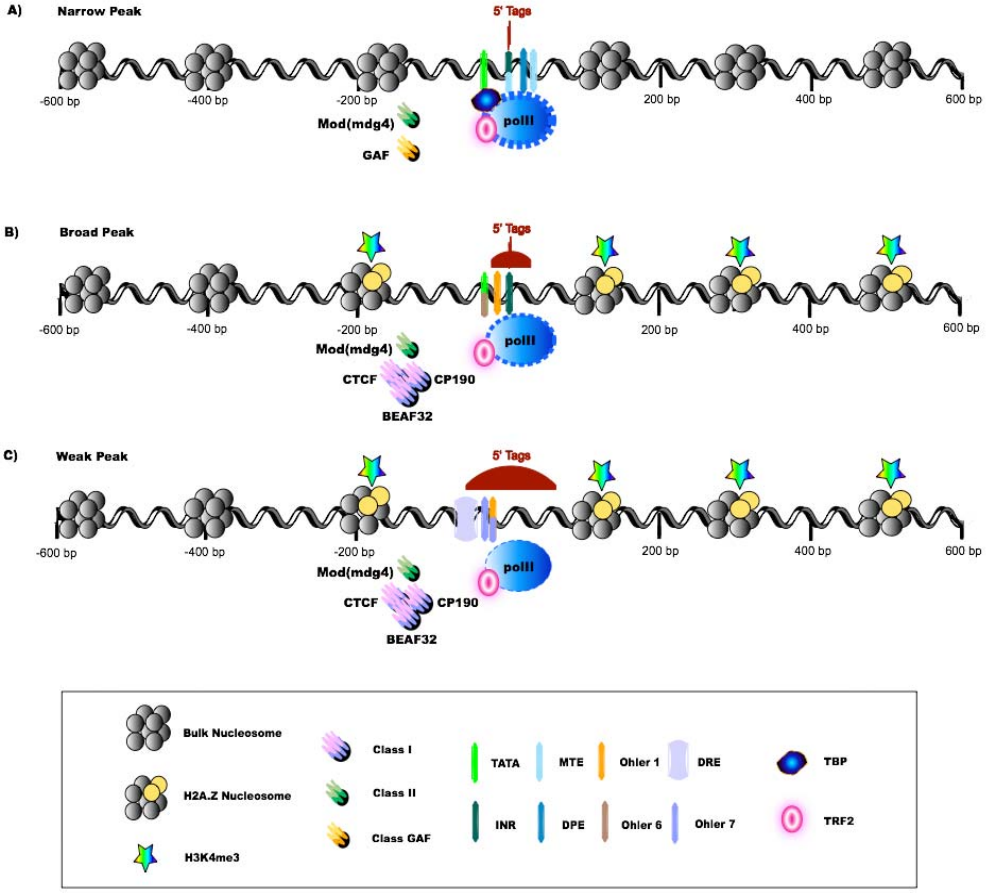


Figure 75: Divergent Strategies for Transcription Initiation

The aggregation of differences in transcription factor binding sites, nucleosome organization, histone variants and chromatin marks, and insulator elements, indicate divergent strategies for transcription initiation in both fruit fly and human. **NP** promoters are marked by precise positioning of transcription initiation, reflected in the presence of location specific core promoter motifs that interact with a canonical TBP-containing basal complex (Ni et al. 2010; Rach et al. 2009). NP promoters show higher levels of pol-II bound to the region around the TSS. They are also associated with specific chromatin remodellers in fly, namely GAF. **WP** promoters are signified by initiation events which spread over a larger genomic span, reflected in the presence of motifs with lower positional enrichment which have been linked to remodeled basal complexes containing TRF2 in fly. They exhibit a well-defined NFR and well-positioned H2A.Z

nucleosomes, as well as associated histone marks such as H3K4 tri-methylations. WP promoters contain an enrichment of fly Class I insulators (CTCF, CP190, BEAF32). The **BP** promoters have a combination of features from both transcriptional programs. While chromatin organization is conserved, some of the known core promoter sequence elements depicted have so far only been found in fly (Ohler 1, DRE, Ohler 6, Ohler 7, MTE). Pol-II and the insulator proteins are depicted at the maximum binding locations; sizes of the transcriptional components are not drawn to scale.

Our findings show that the periodic distribution of nucleosomes in the vicinity of TSSs is strongest for promoters with broad initiation patterns, which have defined NFRs and generally lack the well-spaced canonical core promoter elements. In contrast, the promoters with narrowly defined TSSs exhibit a lower and/or less organized occupancy by nucleosomes, and their precise TSS positions can be largely explained by enriched and well-positioned sequence elements, including the canonical core promoter motifs (Megraw et al. 2009; Ni et al. 2010). In addition, recently defined insulator classes showed distinct associations: Class I insulators (which include CTCF) were associated with H2A.Z organization and H3K4me3 at WP promoters, and class II insulators were evenly distributed. Conversely, GAF and pol-II showed higher levels at NP promoters. The enrichment of the *Drosophila* GAF protein at NP promoters was intriguing, as it is a protein with many reported roles in transcription and chromatin remodeling (Adkins, Hagerman, and Georgel 2006), and may assist transcription initiation at NP promoters in lieu of the lower or less organized nucleosome occupancy. For instance, GAF forms a multimer in replacement of the NFR to establish proper nucleosome organization

(Katsani, Hajibagheri, and Verrijzer 1999), and is enriched at genes with polymerase stalling (Hendrix et al. 2008).

These results agree with previous observations for yeast, where promoters have been divided into two groups based on nucleosome occupancy: Well-defined nucleosome free regions flanked by nucleosomes (Depleted Proximal Nucleosome, DPN) and variable nucleosome positioning without a clear NFR (Occupied Proximal Nucleosome, OPN). While yeast CAGE-like data is not available at a scale needed for the identification and assignment of initiation patterns, NP and WP promoters in fruit fly and human are highly likely correspond to these two classes (Tirosh and Barkai 2008): OPN promoters have a low association with H2A.Z, a high transcriptional plasticity, and are enriched for TATA boxes, while the opposite is true for Depleted Proximal Nucleosome (DPN) promoters. Taken together, our work supports the model established by Barkai et al., in which OPN/NP promoters regulate specific functions in response to specific conditions, while DPN/WP promoters are less variable because they perform housekeeping functions in the cell.

Two different factors may contribute to the divergent features of transcription initiation. First, the chromatin architecture may be fundamentally different between the promoter classes, as illustrated by the nucleosome organization and DNA sequence features. Thus, the group of broad promoters with "less" regulation such as housekeeping genes will have a well-defined NFR accessible to polymerase, in which a

TSS is not well defined and initiation occurs at multiple locations. In NP promoters, TSSs are well defined by sequence elements including canonical core promoter motifs and insulators such as GAF, which actively recruit the polymerase to specific locations instead of defining an overall accessible area. Nucleosome organization in these promoters is consequently less pronounced in average profiles, due to the lack of a common NFR as reflected in the human DHS data. The overall higher pol-II level at the TSS of actively expressed genes with NP promoters also suggests that polymerase stalling is involved as an additional regulatory step important for genes under precise regulation (Nechaev et al. 2009).

Second, the differences in nucleosome and chromatin patterns we observed may result from differences in the pattern and duration of active transcription. It has been suggested that nucleosomes are properly positioned through repeated rounds of active transcription (Henikoff and Ahmad 2005; Zhang et al. 2009). This would support the observation that broad promoters show a greater degree of nucleosome organization within the transcript, and the combinations of histone variants and chromatin marks (such as H2A.Z and H2K4me3) traditionally associated with active transcription, as they are enriched in constitutively expressed genes (Engstrom et al. 2007). In turn, NP promoters are associated with specific time points during embryogenesis (Engstrom et al. 2007), and the lack of constant transcription would lead to a minimal positioning of nucleosomes. Such promoters may have distinct chromatin patterns; for instance, a

higher rate of H3 turnover was observed at OPN promoters in yeast (Tirosh and Barkai 2008), and the presence of GAF has been associated with H3.3 replacement (Mito, Henikoff, and Henikoff 2007), suggesting the possibility that NP promoters may have a higher association with H3.3 replacement. Overall, the question of whether well-positioned nucleosomes help to recruit the transcriptional machinery to the correct location; whether these periodic patterns result from differences in the pattern of active transcription; or rather a combination of both, remains open to further investigation.

As more data becomes available through large-scale efforts such as the modENCODE and ENCODE projects, the divergent strategies of gene regulation will become better characterized throughout development and differentiation in model organisms and human. They may have further implications on epigenetic inheritance, cellular memory, evolvability, and the development of disease (Bernstein, Meissner, and Lander 2007; Tirosh, Barkai, and Verstrepen 2009). Together, these data provide a foundation for deepening our knowledge of the interplay between transcription and epigenetic architecture, and to move our understanding of the genome from a static sequence code to dynamic regulatory networks.

8. Summary and Future Directions

8.1 Major Contributions of This Work

The results in this work have greatly advanced the field of transcription initiation. In spite individual cases, it was once believed that transcription was initiated from a single site, the site was utilized under all spatiotemporal conditions, and each TSS could be identified by a TATA box in the promoter (Schmid et al. 2006). Our work has shown that all of these characteristics are not true in the fruit fly genome.

8.1.1 TSS Identification

Through the identification of TSSs from clustered ESTs, and the experimental PEAT technology, genes with alternative TSS were observed throughout the fruit fly genome. This showed that much in the same way that multiple protein isoforms can be derived from one transcript, so too can multiple TSSs be utilized for one gene. It also suggested that a more complex code exists between the usage of TSSs and the transcriptional selection of downstream sequence.

The ESTs and PEAT technology further showed that instead of transcription being initiated from single 'sites', like the start codon for translation, it begins over varying stretches of DNA that can be characterized by three main patterns: Narrow Peak (NP), Broad Peak (BP), and Weak Peak (WP). Transcription is initiated at NP promoters in focused locations, at WP promoters over larger dispersed regions, and at BP

promoters in focused locations over larger spans of DNA, a combination of NP and WP promoters. In this way, transcription initiation can no longer be thought of as occurring at a single site, but rather at multiple promoter patterns upstream of translated sequence.

8.1.2 Spatiotemporal Utilization

The existence of promoter patterns gave insight into the spatiotemporal expression of genes. The application of Shannon entropy to the clustered ESTs showed that a vast number of promoters were utilized under specific conditions. While instances of individual promoters with utilization in all conditions existed, they occurred less frequently than expected by chance. This specificity of the transcriptional machinery was found from a method that uses conditions with a low resolution, such as tissues like the head and ovary, which demonstrated the sheer importance of the spatiotemporal code in transcription initiation. The libraries of clustered ESTs also revealed that alternative TSSs frequently had different patterns of spatiotemporal expression. This showed that expression profiles should be considered according to individual promoters, rather than by whole gene assignments, as customary.

When the presence of the promoter patterns were evaluated across the temporal stages of fruit fly development, WP promoters had higher utilization during earlier time points, and NP promoters had higher utilization during later time points. This

supported the distinct temporal utilization of the motifs in the promoter patterns, and suggested differences in their functional roles throughout embryogenesis. Together, the spatial and temporal promoter associations showed that TSSs are biologically meaningful signatures of the regulation of expression.

8.1.3 Core Promoter Architecture

The distinct utilization of promoters across spatiotemporal conditions raised questions about the existence of the TATA motif in all promoters. Earlier work had shown that this was not true (Ohler 2006) however, the biological reasoning behind the presence or absence of the TATA was unclear. This provided incentive for us to investigate occurrences of different motifs across the promoter patterns. We divided the eight previously discovered essential core promoter motifs into two categories: those with a position enrichment, and those without one, and compared them to the promoter patterns (Ohler et al. 2002). Cooperative modules of the position-enriched motifs were overrepresented in NP promoters, and modules of the non-position enriched motifs were found in the WP promoters. The BP promoters contained a combination of motifs from both categories. When transcription factors binding profiles were analyzed, higher levels of the position enriched TBP were found at NP promoters, and higher levels of the less-position enriched TRF2 were found at WP promoters. The binding profiles for BP promoters were a mixture of the two.

These associations showed that the core promoter motifs and the transcription factors binding to them were fundamentally different across promoter patterns. This provided a biologically meaningful explanation for the establishment of the promoter patterns from the 5' experimental data, as the genomic span of the TSS data directly resulted from the position enrichment of the motifs in the promoters. The associations also provided insight into the spatiotemporal code, because the distinct spatiotemporal promoter patterns ultimately reflected differences in the utilization of binding sites and the presence of repertoires of transcription factors.

8.1.4 Epigenetic Modifications

In addition to answering important questions about the shape, motif composition, and spatiotemporal regulation of TSSs, our work provided a key link between transcription initiation and the epigenetic architecture of the *Drosophila* genome. When the promoter patterns were compared to the locations of H2A.Z, and H3K4 methylations, WP and BP promoters had a higher association to the organization of these features than NP promoters. H2A.Z and H3K4 methylations have previously been correlated with gene expression (Barski et al. 2007; Mavrich et al. 2008) however, this was the first time that differences in their prevalence were shown to correspond with the transcriptional machinery. These results supported the model in which NP

promoters are determined from their position-enriched motifs, while WP promoters rely more heavily upon nucleosome organization and chromatin structure.

Insulators were also investigated because they disrupt communication between enhancers and promoters, prevent the spread of chromatin, and mark differentially expressed genes (Negre et al. 2010). The results showed the existence of promoter pattern specific insulator classes. This confirmed that not only are localized sequence and epigenetic features indicative of promoter patterns, but their regulators are as well. With these discoveries, TSSs can no longer be considered simple static markers in the genome, but rather dynamic patterns resulting from a combination of histone modifications, chromatin structure, insulator regulation, and localized promoter architecture.

8.2 Cracking the Transcription Initiation Code: Where Do We Go From Here?

Transcription initiation is an exciting field of research that has received much attention in recent years. The results of this work provide a foundation for future explorations in understanding the spatiotemporal code of gene regulation.

8.2.1 Improvement of Experimental and Computational Techniques

One of the first ways that we can increase the accuracy of TSSs is by improving current methods for experimentally capturing 5′ ends. Quality scores assigned to the

ends of sequenced reads used in high throughput technologies can be inconsistent, leading to incorrect TSS calls. In addition, the increasing magnitude of transcript isoforms observed in fruit fly demonstrate the need for a fast and inexpensive method of identifying TSSs simultaneously with their corresponding downstream sequence. While the PEAT methodology provided a high quality mapping of TSSs, associations of the promoters to downstream isoforms were limited.

Computational approaches for predicting TSSs can also be improved. The wealth of information learned about promoters and their epigenetic associations in this thesis should be incorporated into computational models. This would improve the accuracy of promoter prediction, and would serve as a measure of our knowledge of transcription initiation. For certain features, this can easily be accomplished, such as using a GHMM to model the lengths of WP promoters. For others, more sophisticated statistical frameworks are required. For instance, epigenetic modifications such as H3K4me1 can occur at genomic intervals far upstream and downstream of TSSs (Barski et al. 2007). To accurately model H3K4me1 in the context of promoter prediction, higher order dependencies must be incorporated. However, it is often difficult for computational frameworks, such as Markov models, to do so. For this reason, methods, such as conditional random fields (CRFs) (Vinson et al.), that can accommodate high throughput data with long-range dependencies and non-probabilistic values should be explored.

8.2.2 Expansion of Available Data Sets

The transcription factors, histone replacements, and chromatin modifications assessed here are only a subset of all regulatory mechanisms within each cell. Our work shows that Narrow Peak, Broad Peak, and Weak Peak promoters have fundamentally different transcriptional programs however, the complete repertoire of components for each remains to be determined. The modENCODE consortium is an international collaboration of researchers that are currently generating data for numerous transcription factor binding locations, and epigenetic modifications, including histone replacements, and other chromatin modifications in the histone code (modENCODE 2010). This data is an invaluable resource for the determination of additional promoter associations.

The spatiotemporal conditions in which the transcriptional machinery is evaluated should also be expanded upon. Our results have shown that transcription initiation is a dynamic process that changes over time and in different tissues. Often studies compare data derived from non-comparable conditions. For instance, TSSs from hours 0-12 of development may be compared to polII data from Scheider S2 cells that are a specialized cell line derived from embryos during hours 22-24. This may obscure true associations, leading to incorrect conclusions about the transcriptional machinery. To gain a deeper understanding of promoter architecture, it should be studied using corresponding data sets from TSSs and cell types or time periods with a higher

specificity. The modENCODE project is currently generating TSS and regulatory factor data from various stages of *Drosophila* development, and different tissues (modENCODE 2010). This feat was once believed to be too large to accomplish, the advancement of technology has now made it possible.

8.2.3 Applications Across Species and in Diseases

The basic properties of transcription initiation should be explored across species. Current methods for cross species comparisons use the *D.melanogaster* genome as a reference to infer properties in other genomes. The existence of similar transcriptional components and their functional usage in other species remains unknown. The fruit fly is an ideal organism for studying the evolution of the transcriptional machinery because 11 sister genomes have been sequenced (Clark et al. 2007). This is the largest wealth of phylogenetic data to date to evaluate the origin and differentiation of the divergent promoter patterns.

The knowledge of the promoter patterns should also be evaluated in light of diseases. Karen Adelman's lab at the National Institute of Environmental Health Sciences (NIEHS) has shown that the pausing of the polII machinery is associated with Narrow Peak promoters, and the regulation of gene expression in the fruit fly immune system (Adelman et al. 2009). Immune diseases can result from the mis-regulation of

genes, and the knowledge of promoter properties may give insight into the cause of human diseases, such as psoriatic arthritis.

Appendix A Hierarchical Clustering of ESTs and Identified TSS

In Chapter 3, ESTs from the BDGC were aligned to Release 4.3 of the Drosophila genome. The ESTs were first grouped into clusters that were separated by a gap distance, then the clusters were grouped into (sub-)clusters by the standard deviation of the frequencies of ESTs, and lastly the (sub-)clusters were selectively filtered using additional criteria. The frequencies of ESTs in the libraries are given for the initial groupings of ESTs, the (sub-) clusters after clustering, and the TSSs that were chosen from each (sub-)cluster. In addition, the coordinates of all of the isoforms for a gene were compared, and the locations of the most downstream start codon, and the most upstream stop codon are included, along with all of the isoform IDs. Due to the large size of the file, the clustering information for ten genes that had developmental regulation out of the total of 3,990 genes is presented here. The case study gene *tramtrack* (CG1856) is included in this set.

Gene ID
Chromosome #
Transcription_Orientation +/-
Transcript_Isoform_Coordinates start_codon stop_codon transcript_ID
Tags_clustered_by_gap_distance Initial_clusters_are_separated_by_[()]
Corresponding_frequencies_of_gap_distance_clustering
    EST_frequencies_of_clusters
Tags_clustered_by_gap_distance_and_standard_deviation
    Clusters_are_separated_by_[],_sub-clusters_are_separated_by_()
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clustering
    EST_frequencies_of_(sub-)clusters
****
Tags_after_all_clustering_criterion_removal
    (sub)_clusters_remaining_after_all_clustering

Corresponding_frequencies_after_all_cllustering_criterion_removal
EST_frequencies_(sub-)clusters
LIBRARY EST_counts
****
Identified_TSS (sub-)cluster_TSS
Corresponding_TSS_frequencies EST_frequencies_TSS
LIBRARY EST_counts_TSS

Gene CG10334
Chromosome 2L
Transcription_Orientation -
Transcript_Isoform_Coordinates
19564926 19564222 CG10334-PG:CG10334-PE:CG10334-PD:CG10334-PA:CG10334-PF:CG10334-PB:CG10334-PC
Tags_clustered_by_gap_distance
[(19567543,19567607,19567701)][(19568502,19568520,19568536,19568537,19568539)][(19571516)][(19573159,19573172,19573176,19573194,19573196,19573198,19573200,19573201,19573202,19573206,19573208,19573212,19573219,19573221,19573222,19573233,19573239,19573241,19573253)]
Corresponding_frequencies_of_gap_distance_clustering
[(1,1,1)][(1,2,3,2,2)][(3)][(3,1,2,1,1,1,1,1,1,1,1,1,2,3,1,1,1,2)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(19567543)(19567607)(19567701)][(19568502)(19568520,19568536,19568537,19568539)][(19571516)][(19573159,19573172,19573176)(19573194,19573196,19573198,19573200,19573201,19573202,19573206,19573208,19573212)(19573219,19573221,19573222)(19573233,19573239,19573241)(19573253)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clustering
[(1)(1)(1)][(1)(2,3,2,2)][(3)][(3,1,2)(1,1,1,1,1,1,1,1,1)(1,2,3)(1,1,1)(2)]
****
Tags_after_all_clustering_criterion_removal
[(19568520,19568536,19568537,19568539)][(19571516)][(19573159,19573172,19573176)(19573194,19573196,19573198,19573200,19573201,19573202,19573206,19573208,19573212)(19573219,19573221,19573222)(19573233,19573239,19573241)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(2,3,2,2)][(3)][(3,1,2)(1,1,1,1,1,1,1,1,1)(1,2,3)(1,1,1)]
RE_RIKEN_EMBRYO [(0,3,0,0)][(3)][(0,0,2)(1,0,0,0,0,0,0,0,0)(1,0,3)(0,1,0)]
LD_EMBRYO [(0,0,0,0)][(0)][(2,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,1)]
LP_LARVA_PUPA [(0,0,0,0)][(0)][(1,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
RH_RIKEN_HEAD [(0,0,2,2)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
HL_HEAD [(0,0,0,0)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
GH_HEAD [(1,0,0,0)][(0)][(0,1,0)(0,1,0,0,0,1,1,0,0)(0,0,0)(0,0,0)]
GM_OVARY [(0,0,0,0)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
AT_TESTES [(0,0,0,0)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
UT_TESTES [(0,0,0,0)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]

bs_TESTES [(0,0,0,0)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
SD_SCHNEIDER_CELLS [(1,0,0,0)][(0)][(0,0,0)(0,0,1,1,1,0,0,1,1)(0,2,0)(1,0,0)]
EN_MBN2 [(0,0,0,0)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
EC_FAT_BODY [(0,0,0,0)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
OTHERS [(0,0,0,0)][(0)][(0,0,0)(0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0)]
****
Identified_TSS [(19568536)][(19571516)][(19573159)(19573222)]
Corresponding_TSS_frequencies [(3)][(3)][(3)(3)]
RE_RIKEN_EMBRYO [(3)][(3)][(0)(3)]
LD_EMBRYO [(0)][(0)][(2)(0)]
LP_LARVA_PUPA [(0)][(0)][(1)(0)]
RH_RIKEN_HEAD [(0)][(0)][(0)(0)]
HL_HEAD [(0)][(0)][(0)(0)]
GH_HEAD [(0)][(0)][(0)(0)]
GM_OVARY [(0)][(0)][(0)(0)]
AT_TESTES [(0)][(0)][(0)(0)]
UT_TESTES [(0)][(0)][(0)(0)]
bs_TESTES [(0)][(0)][(0)(0)]
SD_SCHNEIDER_CELLS [(0)][(0)][(0)(0)]
EN_MBN2 [(0)][(0)][(0)(0)]
EC_FAT_BODY [(0)][(0)][(0)(0)]
OTHERS [(0)][(0)][(0)(0)]

Gene CG1856
Chromosome 3R
Transcription_Orientation +
Transcript_Isoform_Coordinates
    27553013 27560218 CG1856-PE:CG1856 PB:CG1856-PA
    27553013 27556420 CG1856-PC:CG1856-PF:CG1856-PD
Tags_clustered_by_gap_distance
    [(27539591,27539606)][(27539741,27539771,27539782,27539794)][(27550731,27550733,275507
    37,27550745,27550749,27550754)][(27551503,27551504,27551532,27551542)][(27552853)][(275
    53800,27553833)]
Corresponding_frequencies_of_gap_distance_clustering
    [(2,1)][(2,3,1,1)][(1,8,1,1,1,2)][(3,12,1,1)][(1)][(1,1)]
Tags_clustered_by_gap_distance_and_standard_deviation
    [(27539591,27539606)][(27539741)(27539771,27539782)(27539794)][(27550731,27550733,2755
    0737,27550745,27550749,27550754)][(27551503,27551504)(27551532,27551542)][(27552853)][(
    27553800)(27553833)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clustering
    [(2,1)][(2)(3,1)(1)][(1,8,1,1,1,2)][(3,12)(1,1)][(1)][(1)(1)]
****
Tags_after_all_clustering_criterion_removal
    [(27539591,27539606)][(27539771,27539782)][(27550731,27550733,27550737,27550745,275507
    49,27550754)][(27551503,27551504)]
Corresponding_frequencies_after_all_clustering_criterion_removal
    [(2,1)][(3,1)][(1,8,1,1,1,2)][(3,12)]
RE_RIKEN_EMBRYO [(2,0)][(3,0)][(1,8,1,1,0,0)][(3,12)]

LD_EMBRYO [(0,1)][(0,0)][(0,0,0,0,1,1)][(0,0)]
LP_LARVA_PUPA [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
RH_RIKEN_HEAD [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
HL_HEAD [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
GH_HEAD [(0,0)][(0,1)][(0,0,0,0,0,0)][(0,0)]
GM_OVARY [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
AT_TESTES [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
UT_TESTES [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
bs_TESTES [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
SD_SCHNEIDER_CELLS [(0,0)][(0,0)][(0,0,0,0,0,1)][(0,0)]
EN_MBN2 [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
EC_FAT_BODY [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
OTHERS [(0,0)][(0,0)][(0,0,0,0,0,0)][(0,0)]
****
Identified_TSS [(27539771)][(27550733)][(27551504)]
Corresponding_TSS_frequencies [(3)][(8)][(12)]
RE_RIKEN_EMBRYO [(3)][(8)][(12)]
LD_EMBRYO [(0)][(0)][(0)]
LP_LARVA_PUPA [(0)][(0)][(0)]
RH_RIKEN_HEAD [(0)][(0)][(0)]
HL_HEAD [(0)][(0)][(0)]
GH_HEAD [(0)][(0)][(0)]
GM_OVARY [(0)][(0)][(0)]
AT_TESTES [(0)][(0)][(0)]
UT_TESTES [(0)][(0)][(0)]
bs_TESTES [(0)][(0)][(0)]
SD_SCHNEIDER_CELLS [(0)][(0)][(0)]
EN_MBN2 [(0)][(0)][(0)]
EC_FAT_BODY [(0)][(0)][(0)]
OTHERS [(0)][(0)][(0)]

Gene CG2671
Chromosome 2L
Transcription_Orientation -
Transcript_Isoform_Coordinates
17136 11215 CG2671-PC:CG2671-PA
15648 11215 CG2671-PE:CG2671-PF:CG2671-PD
19944 11215 CG2671-PB
Tags_clustered_by_gap_distance
[(11445,11500)][(11833,11888)][(12148)][(13749,13822)][(17080)][(174
95)][(18473,18491,18522,18534,18536,18537,18541,18548,18550,18560,
18567,18583)][(21200,21285,21309,21327,21357,21369,21372)]
Corresponding_frequencies_of_gap_distance_clustering
[(1,1)][(1,1)][(1)][(1,1)][(1)][(1)][(1,1,1,2,1,1,1,1,1,4,2)][(1,1,2,1,1,1,5)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(11445)(11500)][(11833)(11888)][(12148)][(13749)(13822)][(17080)][(1
7495)][(18473)(18491)(18522)(18534,18536,18537,18541,18548,18550)(
18560,18567)(18583)][(21200)(21285)(21309)(21327)(21357,21369,21372)]

Corresponding_frequencies_of_gap_distance_and_standard_deviation_clusteri
ng
[(1)(1)][(1)(1)][(1)][(1)(1)][(1)][(1)][(1)(1)(1)(2,1,1,1,1,1)(1,4)(2)][(1)(1)(
2)(1)(1,1,5)]
****
Tags_after_all_clustering_criterion_removal
[(18534,18536,18537,18541,18548,18550)(18560,18567)][(21357,21369,
21372)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(2,1,1,1,1,1)(1,4)][(1,1,5)]
RE_RIKEN_EMBRYO [(2,0,0,1,1,0)(0,4)][(0,0,5)]
LD_EMBRYO [(0,1,1,0,0,0)(1,0)][(0,0,0)]
LP_LARVA_PUPA [(0,0,0,0,0,0)(0,0)][(0,0,0)]
RH_RIKEN_HEAD [(0,0,0,0,0,0)(0,0)][(0,0,0)]
HL_HEAD [(0,0,0,0,0,0)(0,0)][(0,0,0)]
GH_HEAD [(0,0,0,0,0,0)(0,0)][(0,0,0)]
GM_OVARY [(0,0,0,0,0,0)(0,0)][(0,0,0)]
AT_TESTES [(0,0,0,0,0,0)(0,0)][(0,0,0)]
UT_TESTES [(0,0,0,0,0,0)(0,0)][(0,0,0)]
bs_TESTES [(0,0,0,0,0,0)(0,0)][(0,0,0)]
SD_SCHNEIDER_CELLS [(0,0,0,0,0,1)(0,0)][(0,0,0)]
EN_MBN2 [(0,0,0,0,0,0)(0,0)][(0,0,0)]
EC_FAT_BODY [(0,0,0,0,0,0)(0,0)][(0,0,0)]
OTHERS [(0,0,0,0,0,0)(0,0)][(1,1,0)]
****
Identified_TSS [(18534)(18567)][(21372)]
Corresponding_TSS_frequencies [(2)(4)][(5)]
RE_RIKEN_EMBRYO [(2)(4)][(5)]
LD_EMBRYO [(0)(0)][(0)]
LP_LARVA_PUPA [(0)(0)][(0)]
RH_RIKEN_HEAD [(0)(0)][(0)]
HL_HEAD [(0)(0)][(0)]
GH_HEAD [(0)(0)][(0)]
GM_OVARY [(0)(0)][(0)]
AT_TESTES [(0)(0)][(0)]
UT_TESTES [(0)(0)][(0)]
bs_TESTES [(0)(0)][(0)]
SD_SCHNEIDER_CELLS [(0)(0)][(0)]
EN_MBN2 [(0)(0)][(0)]
EC_FAT_BODY [(0)(0)][(0)]
OTHERS [(0)(0)][(0)]

Gene CG31243
Chromosome 3R
Transcription_Orientation +
Transcript_Isoform_Coordinates
13792204 13833112 CG31243-PG
13792204 13835833 CG31243-PF:CG31243-PA:CG31243-PB

13792204 13793676 CG31243-PH
13792204 13836858 CG31243-PE
Tags_clustered_by_gap_distance
[(13757595)][(13769829,13769835,13769836)]
Corresponding_frequencies_of_gap_distance_clustering
[(3)][(3,1,1)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(13757595)][(13769829,13769835,13769836)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clustering
[(3)][(3,1,1)]
****
Tags_after_all_clustering_criterion_removal
[(13757595)][(13769829,13769835,13769836)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(3)][(3,1,1)]
RE_RIKEN_EMBRYO [(3)][(3,0,0)]
LD_EMBRYO [(0)][(0,0,0)]
LP_LARVA_PUPA [(0)][(0,0,0)]
RH_RIKEN_HEAD [(0)][(0,0,0)]
HL_HEAD [(0)][(0,0,0)]
GH_HEAD [(0)][(0,0,0)]
GM_OVARY [(0)][(0,0,0)]
AT_TESTES [(0)][(0,0,0)]
UT_TESTES [(0)][(0,0,0)]
bs_TESTES [(0)][(0,0,0)]
SD_SCHNEIDER_CELLS [(0)][(0,0,0)]
EN_MBN2 [(0)][(0,0,0)]
EC_FAT_BODY [(0)][(0,0,0)]
OTHERS [(0)][(0,1,1)]
****
Identified_TSS [(13757595)][(13769829)]
Corresponding_TSS_frequencies [(3)][(3)]
RE_RIKEN_EMBRYO [(3)][(3)]
LD_EMBRYO [(0)][(0)]
LP_LARVA_PUPA [(0)][(0)]
RH_RIKEN_HEAD [(0)][(0)]
HL_HEAD [(0)][(0)]
GH_HEAD [(0)][(0)]
GM_OVARY [(0)][(0)]
AT_TESTES [(0)][(0)]
UT_TESTES [(0)][(0)]
bs_TESTES [(0)][(0)]
SD_SCHNEIDER_CELLS [(0)][(0)]
EN_MBN2 [(0)][(0)]
EC_FAT_BODY [(0)][(0)]
OTHERS [(0)][(0)]

Gene CG3725
Chromosome 2R
Transcription_Orientation -
Transcript_Isoform_Coordinates
19440384 19436831 CG3725-PF:CG3725-PH:CG3725-PE:CG3725-PD:CG3725-PB:CG3725-PC:CG3725-PG
19440384 19434753 CG3725-PA
Tags_clustered_by_gap_distance
[(19436964,19437018)][(19437181)][(19437333)][(19437469,19437498)][(19437813,19437814,19437899,19437936)][(19440090)][(19440399,19440490)][(19441003,19441005,19441011,19441037,19441045,19441048,19441055,19441060,19441077,19441078,19441085,19441090,19441091,19441093,19441097,19441100,19441103,19441112,19441130,19441147)][(19441318,19441332,19441336,19441337,19441358)][(19441551,19441560,19441597,19441683)][(19443144)][(19443253,19443254,19443256,19443260,19443261,19443262,19443263,19443265,19443269,19443272,19443275,19443276,19443279)][(19443480,19443485)]
Corresponding_frequencies_of_gap_distance_clustering
[(1,1)][(1)][(1)][(1,1)][(2,1,1,1)][(1)][(1,1)][(2,1,1,1,2,1,1,7,6,1,1,4,2,1,1,3,1,6,3,1)][(2,3,2,4,3)][(1,1,1,1)][(1)][(4,7,4,3,3,1,3,4,1,2,1,1,7)][(1,1)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(19436964)(19437018)][(19437181)][(19437333)][(19437469)(19437498)][(19437813,19437814)(19437899)(19437936)][(19440090)][(19440399)(19440490)][(19441003,19441005,19441011)(19441037,19441045,19441048,19441055,19441060)(19441077,19441078,19441085,19441090,19441091,19441093,19441097,19441100,19441103)(19441112)(19441130,19441147)][(19441318,19441332,19441336,19441337)(19441358)][(19441551,19441560)(19441597)(19441683)][(19443144)][(19443253,19443254,19443256,19443260,19443261,19443262,19443263,19443265,19443269,19443272,19443275,19443276,19443279)][(19443480,19443485)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clustering
[(1)(1)][(1)][(1)][(1)(1)][(2,1)(1)(1)][(1)][(1)(1)][(2,1,1)(1,2,1,1,7)(6,1,1,4,2,1,1,3,1)(6)(3,1)][(2,3,2,4)(3)][(1,1)(1)(1)][(1)][(4,7,4,3,3,1,3,4,1,2,1,1,7)][(1,1)]
****
Tags_after_all_clustering_criterion_removal
[(19437813,19437814)][(19441003,19441005,19441011)(19441037,19441045,19441048,19441055,19441060)(19441077,19441078,19441085,19441090,19441091,19441093,19441097,19441100,19441103)(19441112)(19441130,19441147)][(19441318,19441332,19441336,19441337)(19441358)][(19443253,19443254,19443256,19443260,19443261,19443262,19443263,19443265,19443269,19443272,19443275,19443276,19443279)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(2,1)][(2,1,1)(1,2,1,1,7)(6,1,1,4,2,1,1,3,1)(6)(3,1)][(2,3,2,4)(3)][(4,7,4,3,3,1,3,4,1,2,1,1,7)]
RE_RIKEN_EMBRYO
[(0,0)][(0,0,0)(0,0,0,0,7)(0,0,1,0,0,1,0,1,1)(1)(3,0)][(0,0,0,4)(3)][(0,0,0,0,0,0,2,0,0,0,0,0,3)]
LD_EMBRYO
[(0,0)][(0,0,0)(0,0,0,0,0)(1,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,0,0)(0)][(0,0,0,0,0,0,0,0,0,1,0,0,0)]
LP_LARVA_PUPA
[(0,0)][(2,1,1)(0,0,0,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(2,0,0,0)(0)][(0,2,1,1,0,0,0,1,0,0,1,0,2)]

RH_RIKEN_HEAD
[(0,0)][(0,0,0)(0,0,0,1,0)(0,0,0,4,0,0,0,0,0)(5)(0,0)][(0,3,1,0)(0)][(4,0,0,0,0,0,0,0,0,0,0,1,2)]
HL_HEAD
[(0,0)][(0,0,0)(0,0,0,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,0,0)(0)][(0,1,0,0,1,0,0,0,1,0,0,0,0)]
GH_HEAD
[(0,0)][(0,0,0)(0,0,0,0,0)(2,1,0,0,0,0,0,2,0)(0)(0,0)][(0,0,0,0)(0)][(0,4,3,2,2,1,1,3,0,1,0,0,0)]
GM_OVARY
[(2,0)][(0,0,0)(0,0,0,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,0,0)(0)][(0,0,0,0,0,0,0,0,0,0,0,0,0)]
AT_TESTES
[(0,0)][(0,0,0)(0,0,1,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,1,0)(0)][(0,0,0,0,0,0,0,0,0,0,0,0,0)]
UT_TESTES
[(0,0)][(0,0,0)(0,0,0,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,0,0)(0)][(0,0,0,0,0,0,0,0,0,0,0,0,0)]
bs_TESTES
[(0,0)][(0,0,0)(0,0,0,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,0,0)(0)][(0,0,0,0,0,0,0,0,0,0,0,0,0)]
SD_SCHNEIDER_CELLS
[(0,0)][(0,0,0)(1,2,0,0,0)(3,0,0,0,2,0,1,0,0)(0)(0,1)][(0,0,0,0)(0)][(0,0,0,0,0,0,0,0,0,0,0,0,0)]
EN_MBN2
[(0,0)][(0,0,0)(0,0,0,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,0,0)(0)][(0,0,0,0,0,0,0,0,0,0,0,0,0)]
EC_FAT_BODY
[(0,0)][(0,0,0)(0,0,0,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,0,0)(0)][(0,0,0,0,0,0,0,0,0,0,0,0,0)]
OTHERS
[(0,1)][(0,0,0)(0,0,0,0,0)(0,0,0,0,0,0,0,0,0)(0)(0,0)][(0,0,0,0)(0)][(0,0,0,0,0,0,0,0,0,0,0,0,0)]
****
Identified_TSS
[(19437813)][(19441060)(19441112)][(19441337)(19441358)][(19443279)]
Corresponding_TSS_frequencies [(2)][(7)(6)][(4)(3)][(7)]
RE_RIKEN_EMBRYO [(0)][(7)(1)][(4)(3)][(3)]
LD_EMBRYO [(0)][(0)(0)][(0)(0)][(0)]
LP_LARVA_PUPA [(0)][(0)(0)][(0)(0)][(2)]
RH_RIKEN_HEAD [(0)][(0)(5)][(0)(0)][(2)]
HL_HEAD [(0)][(0)(0)][(0)(0)][(0)]
GH_HEAD [(0)][(0)(0)][(0)(0)][(0)]
GM_OVARY [(2)][(0)(0)][(0)(0)][(0)]
AT_TESTES [(0)][(0)(0)][(0)(0)][(0)]
UT_TESTES [(0)][(0)(0)][(0)(0)][(0)]
bs_TESTES [(0)][(0)(0)][(0)(0)][(0)]
SD_SCHNEIDER_CELLS [(0)][(0)(0)][(0)(0)][(0)]
EN_MBN2 [(0)][(0)(0)][(0)(0)][(0)]
EC_FAT_BODY [(0)][(0)(0)][(0)(0)][(0)]
OTHERS [(0)][(0)(0)][(0)(0)][(0)]

Gene CG4898
Chromosome 3R
Transcription_Orientation +
Transcript_Isoform_Coordinates
11110371 11133397 CG4898-PE:CG4898-PJ:CG4898-PG:CG4898-PD:CG4898-PB
11117104 11129445 CG4898-PA

11110371 11131633 CG4898-PF
11110371 11129445 CG4898-PL
11114319 11122661 CG4898-PC:CG4898-PI
11110371 11132847 CG4898-PK
11113177 11122661 CG4898-PH
Tags_clustered_by_gap_distance
[(11107254,11107274,11107275,11107277,11107278,11107283,11107287,11107288,11107289,11107292,11107293,11107298,11107307,11107326)][(11110327,11110333)][(11110448)][(11111000)][(11112141)][(11113755)][(11115090)][(11116668,11116674,11116678,11116686,11116687,11116688,11116703)][(11120837)][(11121179,11121204)][(11125826,11125861)][(11126765)][(11127124)][(11128895)]
Corresponding_frequencies_of_gap_distance_clustering
[(1,76,2,1,5,2,1,2,1,2,3,1,1,2)][(1,2)][(2)][(1)][(1)][(1)][(1)][(3,2,1,1,2,1,1)][(1)][(1,1)][(1,1)][(1)][(1)][(1)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(11107254,11107274,11107275,11107277,11107278,11107283,11107287,11107288,11107289,11107292,11107293,11107298,11107307,11107326)][(11110327,11110333)][(11110448)][(11111000)][(11112141)][(11113755)][(11115090)][(11116668,11116674,11116678,11116686,11116687,11116688)(11116703)][(11120837)][(11121179)(11121204)][(11125826)(11125861)][(11126765)][(11127124)][(11128895)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clusteri
ng
[(1,76,2,1,5,2,1,2,1,2,3,1,1,2)][(1,2)][(2)][(1)][(1)][(1)][(1)][(3,2,1,1,2,1)(1)][(1)][(1)(1)][(1)(1)][(1)][(1)]
****
Tags_after_all_clustering_criterion_removal
[(11107254,11107274,11107275,11107277,11107278,11107283,11107287,11107288,11107289,11107292,11107293,11107298,11107307,11107326)][(11110327,11110333)][(11116668,11116674,11116678,11116686,11116687,11116688)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(1,76,2,1,5,2,1,2,1,2,3,1,1,2)][(1,2)][(3,2,1,1,2,1)]
RE_RIKEN_EMBRYO [(0,76,2,1,5,0,0,0,0,1,0,0,0,2)][(1,0)][(3,2,1,0,0,0)]
LD_EMBRYO [(0,0,0,0,0,0,0,0,1,0,1,0,0,0)][(0,0)][(0,0,0,1,0,0)]
LP_LARVA_PUPA [(0,0,0,0,0,1,0,1,0,0,2,0,0,0)][(0,1)][(0,0,0,0,0,0)]
RH_RIKEN_HEAD [(1,0,0,0,0,0,0,0,0,0,0,0,0,0)][(0,0)][(0,0,0,0,0,0)]
HL_HEAD [(0,0,0,0,0,0,0,0,0,0,0,0,1,0)][(0,0)][(0,0,0,0,0,0)]
GH_HEAD [(0,0,0,0,0,0,0,1,1,0,0,0,0,0,0)][(0,1)][(0,0,0,0,1,0)]
GM_OVARY [(0,0,0,0,0,0,1,0,0,0,0,0,0,0,0)][(0,0)][(0,0,0,0,0,0)]
AT_TESTES [(0,0,0,0,0,0,0,0,0,0,0,1,0,0)][(0,0)][(0,0,0,0,0,0)]
UT_TESTES [(0,0,0,0,0,0,0,0,0,0,0,0,0,0)][(0,0)][(0,0,0,0,0,0)]
bs_TESTES [(0,0,0,0,0,0,0,0,0,1,0,0,0,0)][(0,0)][(0,0,0,0,0,0)]
SD_SCHNEIDER_CELLS [(0,0,0,0,0,0,0,0,0,0,0,0,0,0)][(0,0)][(0,0,0,0,1,1)]
EN_MBN2 [(0,0,0,0,0,0,0,0,0,0,0,0,0,0)][(0,0)][(0,0,0,0,0,0)]
EC_FAT_BODY [(0,0,0,0,0,0,0,0,0,0,0,0,0,0)][(0,0)][(0,0,0,0,0,0)]
OTHERS [(0,0,0,0,0,0,0,0,0,0,0,0,0,0)][(0,0)][(0,0,0,0,0,0)]
****
Identified_TSS [(11107274)][(11116668)]
Corresponding_TSS_frequencies [(76)][(3)]
RE_RIKEN_EMBRYO [(76)][(3)]

LD_EMBRYO [(0)][(0)]
LP_LARVA_PUPA [(0)][(0)]
RH_RIKEN_HEAD [(0)][(0)]
HL_HEAD [(0)][(0)]
GH_HEAD [(0)][(0)]
GM_OVARY [(0)][(0)]
AT_TESTES [(0)][(0)]
UT_TESTES [(0)][(0)]
bs_TESTES [(0)][(0)]
SD_SCHNEIDER_CELLS [(0)][(0)]
EN_MBN2 [(0)][(0)]
EC_FAT_BODY [(0)][(0)]
OTHERS [(0)][(0)]

Gene CG8989
Chromosome X
Transcription_Orientation -
Transcript_Isoform_Coordinates
8999882 8999333 CG8989-PA:CG8989-PC:CG8989-PB
Tags_clustered_by_gap_distance
[(8998683)][(8999674,8999734,8999753,8999760,8999846,8999908)][(9000223,9000226)][(9000834)][(9001846,9001852,9001853,9001857,9001858,9001860,9001861,9001868,9001870,9001884,9001885,9001886,9001888,9001905)]
Corresponding_frequencies_of_gap_distance_clustering
[(1)][(1,1,1,1,1,1)][(5,3)][(1)][(2,1,2,1,1,1,2,2,1,1,2,30,1,1)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(8998683)][(8999674)(8999734)(8999753,8999760)(8999846)(8999908)][(9000223,9000226)][(9000834)][(9001846,9001852,9001853,9001857,9001858,9001860,9001861,9001868,9001870)(9001884,9001885,9001886,9001888)(9001905)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clustering
[(1)][(1)(1)(1,1)(1)(1)][(5,3)][(1)][(2,1,2,1,1,1,2,2,1)(1,2,30,1)(1)]
****
Tags_after_all_clustering_criterion_removal
[(9000223,9000226)][(9001846,9001852,9001853,9001857,9001858,9001860,9001861,9001868,9001870)(9001884,9001885,9001886,9001888)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(5,3)][(2,1,2,1,1,1,2,2,1)(1,2,30,1)]
RE_RIKEN_EMBRYO [(5,0)][(0,0,0,0,0,0,0,1,1)(1,1,8,1)]
LD_EMBRYO [(0,0)][(2,1,1,1,0,1,2,0,0)(0,0,0,0)]
LP_LARVA_PUPA [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
RH_RIKEN_HEAD [(0,3)][(0,0,0,0,0,0,0,1,0)(0,1,22,0)]
HL_HEAD [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
GH_HEAD [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
GM_OVARY [(0,0)][(0,0,1,0,0,0,0,0,0)(0,0,0,0)]
AT_TESTES [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
UT_TESTES [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
bs_TESTES [(0,0)][(0,0,0,0,1,0,0,0,0)(0,0,0,0)]

SD_SCHNEIDER_CELLS [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
EN_MBN2 [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
EC_FAT_BODY [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
OTHERS [(0,0)][(0,0,0,0,0,0,0,0,0)(0,0,0,0)]
****
Identified_TSS [(9000223)][(9001886)]
Corresponding_TSS_frequencies [(5)][(30)]
RE_RIKEN_EMBRYO [(5)][(8)]
LD_EMBRYO [(0)][(0)]
LP_LARVA_PUPA [(0)][(0)]
RH_RIKEN_HEAD [(0)][(22)]
HL_HEAD [(0)][(0)]
GH_HEAD [(0)][(0)]
GM_OVARY [(0)][(0)]
AT_TESTES [(0)][(0)]
UT_TESTES [(0)][(0)]
bs_TESTES [(0)][(0)]
SD_SCHNEIDER_CELLS [(0)][(0)]
EN_MBN2 [(0)][(0)]
EC_FAT_BODY [(0)][(0)]
OTHERS [(0)][(0)]

Gene CG9075
Chromosome 2L
Transcription_Orientation +
Transcript_Isoform_Coordinates
5981834 5985170 CG9075-PA:CG9075-PC
5982820 5985170 CG9075-PB:CG9075-PD
Tags_clustered_by_gap_distance
[(5981759,5981764,5981787,5981788,5981789,5981790,5981791,5981793,5981794,5981795,5981796,5981803,5981812,5981814,5981816,5981819,5981822,5981823)][(5982586,5982600,5982601,5982615,5982617,5982619,5982620,5982622,5982627,5982629,5982634,5982638,5982649,5982652,5982670,5982678,5982708,5982729,5982737,5982743,5982752,5982753,5982754,5982757,5982759,5982763,5982771,5982772)][(5983006,5983012,5983015,5983057,5983082,5983108,5983125,5983179,5983180,5983182)]
Corresponding_frequencies_of_gap_distance_clustering
[(1,1,13,9,9,12,1,9,2,1,11,5,1,6,2,2,1,2)][(1,5,2,1,4,1,2,3,5,4,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)][(1,1,1,2,1,1,3,1,2)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(5981759,5981764)(5981787,5981788,5981789,5981790,5981791,5981793,5981794,5981795,5981796,5981803)(5981812,5981814,5981816,5981819,5981822,5981823)][(5982586)(5982600,5982601)(5982615,5982617,5982619,5982620,5982622,5982627,5982629,5982634,5982638,5982649,5982652)(5982670,5982678)(5982708)(5982729,5982737,5982743)(5982752,5982753,5982754,5982757,5982759,5982763,5982771,5982772)][(5983006,5983012,5983015)(5983057)(5983082)(5983108)(5983125)(5983179,5983180,5983182)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clustering
[(1,1)(13,9,9,12,1,9,2,1,11,5)(1,6,2,2,1,2)][(1)(5,2)(1,4,1,2,3,5,4,1,1,1,1)(1,1)(1)(1,1,1)(1,1,1,1,1,1,1,1,1)][(1,1,1)(1)(2)(1)(1)(3,1,2)]

****
Tags_after_all_clustering_criterion_removal
[(5981787,5981788,5981789,5981790,5981791,5981793,5981794,5981795,5981796,5981803)(5981812,5981814,5981816,5981819,5981822,5981823)][(5982600,5982601)(5982615,5982617,5982619,5982620,5982622,5982627,5982629,5982634,5982638,5982649,5982652)(5982729,5982737,5982743)(5982752,5982753,5982754,5982757,5982759,5982763,5982771,5982772)][(5983006,5983012,5983015)(5983179,5983180,5983182)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(13,9,9,12,1,9,2,1,11,5)(1,6,2,2,1,2)][(5,2)(1,4,1,2,3,5,4,1,1,1,1)(1,1,1)(1,1,1,1,1,1,1,1)][(1,1,1)(3,1,2)]
RE_RIKEN_EMBRYO
[(0,0,0,0,0,0,0,0,0,0)(0,0,0,0,0,0)][(0,2)(0,0,0,0,0,0,0,0,0,0)(0,0,1)(0,0,0,0,0,0,0,0)][(0,0,0)(0,0,0)]
LD_EMBRYO
[(0,0,0,0,0,0,0,0,6,5)(0,4,1,2,0,0)][(0,0)(0,2,1,0,2,0,4,1,1,0,0)(0,1,0)(0,0,0,0,0,0,0,0)][(0,0,0)(3,1,1)]
LP_LARVA_PUPA
[(0,0,0,0,0,0,0,0,0,0)(0,0,0,0,0,0)][(0,0)(1,1,0,2,0,1,0,0,0,0,1)(0,0,0)(0,0,1,0,0,0,0,0)][(0,0,0)(0,0,0)]
RH_RIKEN_HEAD
[(0,0,0,0,0,0,0,0,0,0)(0,0,0,0,0,0)][(5,0)(0,0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0,0,0,0,0,0)][(0,0,0)(0,0,0)]
HL_HEAD
[(0,0,0,0,0,0,0,0,0,0)(0,0,0,0,1,0)][(0,0)(0,0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0,0,0,0,0,0)][(0,0,0)(0,0,0)]
GH_HEAD
[(0,0,0,0,0,0,0,0,0,0)(0,0,1,0,0,0)][(0,0)(0,1,0,0,1,2,0,0,0,0,0)(0,0,0)(0,0,0,
0,0,0,0,0)][(0,0,0)(0,0,1)]
GM_OVARY
[(3,2,7,1,0,5,0,0,2,0)(0,0,0,0,0,0)][(0,0)(0,0,0,0,0,0,0,0,0,0,0)(1,0,0)(1,0,0,0,0,0,0,0)][(0,0,0)(0,0,0)]
AT_TESTES
[(2,0,1,1,0,0,0,1,0,0)(0,0,0,0,0,0)][(0,0)(0,0,0,0,0,0,0,0,0,0,0)(0,0,0)(0,1,0,1,1,1,1,0)][(0,0,0)(0,0,0)]
UT_TESTES
[(0,0,0,0,0,0,0,0,0,0)(0,0,0,0,0,0)][(0,0)(0,0,0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0,0,0,0,0,0)][(0,0,0)(0,0,0)]
bs_TESTES
[(0,0,0,0,0,0,0,0,0,0)(0,0,0,0,0,0)][(0,0)(0,0,0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0,0,0,0,0,0)][(0,0,0)(0,0,0)]
SD_SCHNEIDER_CELLS
[(7,7,1,10,1,4,2,0,3,0)(1,0,0,0,0,0)][(0,0)(0,0,0,0,0,1,0,0,0,1,0)(0,0,0)(0,0,0,0,0,0,0,1)][(0,1,1)(0,0,0)]
EN_MBN2
[(1,0,0,0,0,0,0,0,0,0)(0,2,0,0,0,2)][(0,0)(0,0,0,0,0,1,0,0,0,0,0)(0,0,0)(0,0,0,0,0,0,0,0)][(1,0,0)(0,0,0)]
EC_FAT_BODY
[(0,0,0,0,0,0,0,0,0,0)(0,0,0,0,0,0)][(0,0)(0,0,0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0,0,0,0,0,0)][(0,0,0)(0,0,0)]
OTHERS
[(0,0,0,0,0,0,0,0,0,0)(0,0,0,0,0,0)][(0,0)(0,0,0,0,0,0,0,0,0,0,0)(0,0,0)(0,0,0,0,0,0,0,0)][(0,0,0)(0,0,0)]
****
Identified_TSS [(5981787)][(5982600)]
Corresponding_TSS_frequencies [(13)][(5)]
RE_RIKEN_EMBRYO [(0)][(0)]
LD_EMBRYO [(0)][(0)]
LP_LARVA_PUPA [(0)][(0)]
RH_RIKEN_HEAD [(0)][(5)]
HL_HEAD [(0)][(0)]
GH_HEAD [(0)][(0)]

GM_OVARY [(3)][(0)]
AT_TESTES [(2)][(0)]
UT_TESTES [(0)][(0)]
bs_TESTES [(0)][(0)]
SD_SCHNEIDER_CELLS [(7)][(0)]
EN_MBN2 [(1)][(0)]
EC_FAT_BODY [(0)][(0)]
OTHERS [(0)][(0)]

Gene CG9261
Chromosome 2L
Transcription_Orientation -
Transcript_Isoform_Coordinates
6796481 6790639 CG9261-PC:CG9261-PF
6793369 6790639 CG9261-PD:CG9261-PE:CG9261-PA
Tags_clustered_by_gap_distance
[(6790380)][(6790512,6790516,6790542,6790581,6790587,6790600,6790674)][(6790854)][(6791110,6791143,6791184,6791239)][(6791553)][(6793381)][(6794396)][(6794517,6794594,6794595,6794596,6794597)][(6796445)][(6796765,6796768,6796770,6796776,6796779,6796780)][(6798520)][(6798864)]
Corresponding_frequencies_of_gap_distance_clustering
[(1)][(1,2,1,1,1,1,1)][(1)][(1,2,2,1)][(1)][(1)][(1)][(1,2,20,3,3)][(1)][(2,2,1,3,7,5)][(1)][(1)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(6790380)][(6790512,6790516)(6790542)(6790581,6790587,6790600)(6790674)][(6790854)][(6791110)(6791143)(6791184)(6791239)][(6791553)][(6793381)][(6794396)][(6794517)(6794594,6794595,6794596,6794597)][(6796445)][(6796765,6796768,6796770,6796776,6796779,6796780)][(6798520)][(6798864)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clusteri
ng
[(1)][(1,2)(1)(1,1,1)(1)][(1)][(1)(2)(2)(1)][(1)][(1)][(1)][(1)(2,20,3,3)][(1)][(2,2,1,3,7,5)][(1)][(1)]
****
Tags_after_all_clustering_criterion_removal
[(6790512,6790516)(6790581,6790587,6790600)][(6794594,6794595,6794596,6794597)][(6796765,6796768,6796770,6796776,6796779,6796780)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(1,2)(1,1,1)][(2,20,3,3)][(2,2,1,3,7,5)]
RE_RIKEN_EMBRYO [(0,0)(0,0,0)][(2,17,3,0)][(0,0,0,3,3,5)]
LD_EMBRYO [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
LP_LARVA_PUPA [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
RH_RIKEN_HEAD [(0,0)(0,0,0)][(0,3,0,3)][(0,0,0,0,4,0)]
HL_HEAD [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
GH_HEAD [(0,0)(0,0,0)][(0,0,0,0)][(2,2,1,0,0,0)]
GM_OVARY [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
AT_TESTES [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
UT_TESTES [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
bs_TESTES [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
SD_SCHNEIDER_CELLS [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
EN_MBN2 [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]

EC_FAT_BODY [(0,0)(0,0,0)][(0,0,0,0)][(0,0,0,0,0,0)]
OTHERS [(1,2)(1,1,1)][(0,0,0,0)][(0,0,0,0,0,0)]
****
Identified_TSS [(6794595)][(6796779)]
Corresponding_TSS_frequencies [(20)][(7)]
RE_RIKEN_EMBRYO [(17)][(3)]
LD_EMBRYO [(0)][(0)]
LP_LARVA_PUPA [(0)][(0)]
RH_RIKEN_HEAD [(3)][(4)]
HL_HEAD [(0)][(0)]
GH_HEAD [(0)][(0)]
GM_OVARY [(0)][(0)]
AT_TESTES [(0)][(0)]
UT_TESTES [(0)][(0)]
bs_TESTES [(0)][(0)]
SD_SCHNEIDER_CELLS [(0)][(0)]
EN_MBN2 [(0)][(0)]
EC_FAT_BODY [(0)][(0)]
OTHERS [(0)][(0)]

Gene CG9553
Chromosome 2L
Transcription_Orientation -
Transcript_Isoform_Coordinates
5979522 5973577 CG9553-PA:CG9553-PB:CG9553-PC:CG9553-PD
Tags_clustered_by_gap_distance
[(5974081)][(5976719,5976728,5976774)][(5979469,5979515,5979590,5979625,5979627,5979644,5979661,5979663,5979685,5979693,5979697,5979701,5979707,5979708)][(5980010,5980054,5980058,5980062,5980064,5980065,5980070,5980073,5980074,5980076,5980078,5980079,5980082,5980096,5980097,5980098)][(5980476)][(5980968,5980976,5980979,5980982,5980991,5980995,5981004,5981012,5981017,5981023)]
Corresponding_frequencies_of_gap_distance_clustering
[(1)][(1,1,1)][(1,1,1,1,1,1,1,1,1,2,1,3,1)][(1,1,1,1,1,1,1,4,6,2,3,1,8,1,7,23)][(1)][(3,1,1,28,1,1,1,1,1)]
Tags_clustered_by_gap_distance_and_standard_deviation
[(5974081)][(5976719,5976728)(5976774)][(5979469)(5979515)(5979590)(5979625,5979627)(5979644)(5979661,5979663)(5979685,5979693,5979697,5979701,5979707,5979708)][(5980010)(5980054,5980058,5980062,5980064,5980065,5980070,5980073,5980074,5980076,5980078,5980079,5980082)(5980096,5980097,5980098)][(5980476)][(5980968,5980976,5980979,5980982)(5980991,5980995)(5981004,5981012,5981017,5981023)]
Corresponding_frequencies_of_gap_distance_and_standard_deviation_clusteri
ng
[(1)][(1,1)(1)][(1)(1)(1)(1,1)(1)(1,1)(1,1,2,1,3,1)][(1)(1,1,1,1,1,1,4,6,2,3,1,8)(1,7,23)][(1)][(3,1,1,28)(1,1)(1,1,1,1)]
****
Tags_after_all_clustering_criterion_removal

[(5979685,5979693,5979697,5979701,5979707,5979708)][(5980054,5980058,5980062,5980064,5980065,5980070,5980073,5980074,5980076,5980078,5980079,5980082)(5980096,5980097,5980098)][(5980968,5980976,5980979,5980982)(5981004,5981012,5981017,5981023)]
Corresponding_frequencies_after_all_clustering_criterion_removal
[(1,1,2,1,3,1)][(1,1,1,1,1,1,4,6,2,3,1,8)(1,7,23)][(3,1,1,28)(1,1,1,1)]
RE_RIKEN_EMBRYO
[(0,0,0,0,0,0)][(0,0,0,0,0,0,2,0,0,0,0,0)(1,0,8)][(0,0,1,17)(0,1,1,0)]
LD_EMBRYO [(0,0,2,0,0,1)][(0,1,0,0,0,0,0,1,0,0,0,0)(0,0,0)][(1,0,0,0)(0,0,0,1)]
LP_LARVA_PUPA
[(0,0,0,0,0,0)][(0,0,0,1,0,0,0,3,1,1,0,0)(0,0,0)][(0,0,0,0)(0,0,0,0)]
RH_RIKEN_HEAD
[(0,0,0,0,0,0)][(0,0,0,0,0,0,0,0,0,0,0,0)(0,7,13)][(0,0,0,10)(0,0,0,0)]
HL_HEAD [(0,0,0,0,0,0)][(0,0,0,0,0,0,0,0,0,0,0,0)(0,0,0)][(0,0,0,0)(0,0,0,0)]
GH_HEAD [(0,0,0,0,2,0)][(0,0,1,0,0,0,0,0,0,0,1,1)(0,0,0)][(1,0,0,0)(0,0,0,0)]
GM_OVARY [(0,0,0,0,1,0)][(0,0,0,0,0,0,0,0,0,0,0,0)(0,0,0)][(0,0,0,0)(0,0,0,0)]
AT_TESTES [(0,0,0,0,0,0)][(0,0,0,0,0,0,0,0,0,0,0,0)(0,0,0)][(0,0,0,0)(0,0,0,0)]
UT_TESTES [(0,0,0,0,0,0)][(0,0,0,0,0,0,0,0,0,0,0,0)(0,0,0)][(0,0,0,0)(0,0,0,0)]
bs_TESTES [(0,0,0,0,0,0)][(0,0,0,0,0,0,0,0,0,0,0,0)(0,0,0)][(0,0,0,0)(0,0,0,0)]
SD_SCHNEIDER_CELLS
[(0,0,0,0,0,0)][(0,0,0,0,0,0,1,0,0,0,0,0)(0,0,0)][(1,0,0,0)(0,0,0,0)]
EN_MBN2 [(0,1,0,1,0,0)][(1,0,0,0,1,1,1,2,1,2,0,5)(0,0,0)][(0,0,0,0)(0,0,0,0)]
EC_FAT_BODY [(0,0,0,0,0,0)][(0,0,0,0,0,0,0,0,0,0,0,2)(0,0,0)][(0,0,0,0)(1,0,0,0)]
OTHERS [(1,0,0,0,0,0)][(0,0,0,0,0,0,0,0,0,0,0,0)(0,0,2)][(0,1,0,1)(0,0,0,0)]
****
Identified_TSS [(5980098)][(5980982)]
Corresponding_TSS_frequencies [(23)][(28)]
RE_RIKEN_EMBRYO [(8)][(17)]
LD_EMBRYO [(0)][(0)]
LP_LARVA_PUPA [(0)][(0)]
RH_RIKEN_HEAD [(13)][(10)]
HL_HEAD [(0)][(0)]
GH_HEAD [(0)][(0)]
GM_OVARY [(0)][(0)]
AT_TESTES [(0)][(0)]
UT_TESTES [(0)][(0)]
bs_TESTES [(0)][(0)]
SD_SCHNEIDER_CELLS [(0)][(0)]
EN_MBN2 [(0)][(0)]
EC_FAT_BODY [(0)][(0)]
OTHERS [(2)][(1)]

Appendix B Core Promoter Element Matches

The file contains the position weight matrices and the background models used in Chapters 3 and 4. They were used in the program PATSER to search for motifs in the core promoters of the most 5′ sites in Flybase, the identified TSSs, and the random intergenic sites. Motif matches in the promoter regions of TSSs and TSS cluster groups are listed for ten genes, including those located outside of the preferred motif windows. TSSs without at least one motif match in its promoter are excluded from the file. (m1 = Ohler 1, m2 = DRE, m3 = TATA, m4 = INR, m6 = Ohler 6, m7 = Ohler 7, m9 = DPE, m10 = MTE)

Position Weight Matrices Used in PATSER Searches as Generated In (Ohler et al. 2002) The DPE and MTE are shortened from their original length.

Ohler 1

A	C	G	T
84	75	105	47
111	101	73	26
76	68	42	125
138	76	63	34
4	173	7	127
0	0	309	2
0	0	305	6
0	54	14	243
40	268	1	2
296	4	7	4
0	284	0	27
268	9	9	25
0	280	0	31
9	38	3	261
68	13	191	39

DRE

A	C	G	T
137	35	52	53
78	52	104	43
73	108	24	72
61	32	1	183
263	0	5	9
0	0	0	277
0	277	0	0
0	0	277	0
277	0	0	0
20	0	0	257
196	0	24	57
96	35	85	61
49	124	53	51
94	68	38	77

TATA

A	C	G	T
45	49	89	68
47	59	110	35
33	95	114	9
5	5	0	241
251	0	0	0
0	0	0	251
246	0	0	5
195	0	0	56
251	0	0	0
206	0	0	45
68	24	131	28
27	98	98	28
52	91	65	43
62	67	89	33
57	71	80	43

INR

A	C	G	T
85	56	33	195
0	0	5	364
0	330	2	37
369	0	0	0
0	0	339	30
0	0	0	369
10	98	1	260
24	63	164	118
76	119	95	79
102	88	70	109
89	49	102	129
63	23	72	211
20	114	31	204
25	140	93	111
108	34	185	42

Ohler 6

A	C	G	T
30	14	10	53
8	11	30	58
6	17	17	67
2	31	3	71
23	0	83	1
0	3	101	3
3	6	0	98
106	1	0	0
0	0	0	107
25	0	0	82
7	0	0	100
1	0	1	105
11	17	6	73
18	40	12	37
46	23	31	7

Ohler 7

A	C	G	T
40	18	27	112
15	49	48	85
1	17	50	129
15	44	78	60
28	83	58	28
0	197	0	0
197	0	0	0
26	21	54	96
0	197	0	0
54	52	19	72
0	197	0	0
0	23	0	174
133	4	57	3
60	18	86	33
29	85	9	74

DPE

A	C	G	T
6	0	32	18
3	52	0	1
9	0	44	3
1	3	47	5
8	0	0	48
0	4	0	52
0	31	23	2
6	0	32	18

MTE

A	C	G	T
0	38	2	0
6	8	26	0
40	0	0	0
28	0	8	4
0	40	0	0
0	9	30	1
7	12	21	0
34	0	6	0

Background models constructed from local sequence

Highly Utilized Sites Identified from the most frequent 5' EST ends
A 0.263
C 0.233
G 0.228
T 0.276

Most 5' in Flybase
A 0.287
C 0.220
G 0.222
T 0.271

Random Intergenic Sites
A 0.296
C 0.205
G 0.203
T 0.296

Chr#_CGID_Orientation_InitiationSite motif#(Ohler Gen Biol 2002):location_starting_from_60upstream_of_the_TSS
For instance, m4:57 denotes the INR located at 57-60 = -3 from the TSS, or 3bp upstream of the TSS.

2L_CG10334_-_19568536 m9:86
2L_CG10334_-_19573159_19573222 m1:46::m4:96::m7:24::m7:48::m10:146
3R_CG1856_+_27539771 m4:57::m7:40::m9:58::m9:71
3R_CG1856_+_27550733 m9:35::m10:32
3R_CG1856_+_27551504 m4:22::m4:57::m9:8
2L_CG2671_-_18534_18567 m1:50::m2:35::m10:72
2L_CG2671_-_21372 m1:50::m2:35::m10:72
3R_CG31243_+_13757595 m9:65
3R_CG31243_+_13769829 m4:57::m4:64::m9:40::m10:79
2R_CG3725_-_19441060_19441112 m6:106
2R_CG3725_-_19441337_19441358 m2:57::m4:41::m4:54::m9:42::m10:37
2R_CG3725_-_19443279 m4:57
3R_CG4898_+_11107274 m4:57
3R_CG4898_+_11116668 m4:57::m9:1
X_CG8989_-_9000223 m2:25::m4:58::m7:13::m9:59::m9:73::m10:42
X_CG8989_-_9001886 m7:28
2L_CG9075_+_5981787 m3:75::m7:6
2L_CG9075_+_5982600 m2:42
2L_CG9261_-_6794595 m4:58
2L_CG9553_-_5980098 m4:58
2L_CG9553_-_5980982 m1:52::m6:61::m7:32::m10:14

Appendix C TSS EST Condition Associations

In Chapter 4, Shannon entropy was applied to the frequencies of ESTs located in the (sub-)clusters from which the TSSs were identified. The entropy values were used to classify the utilization of the TSS as condition specific, condition supported, or mixed. The gene, chromosome, orientation, and condition associations for the TSSs of the ten genes enclosed in Appendices A and B are listed.

Gene: CG_ID
Chromosome: #
Orientation: +/-
TSS: Location Pattern: Association

Gene: CG10334
Chromosome: 2L
Orientation: -
TSS: 19568536 Pattern: Embryo_supported,Head_supported
TSS: 19571516 Pattern: Embryo_specific
TSS: 19573159 Pattern: Embryo_supported
TSS: 19573222 Pattern: Embryo_supported

Gene: CG1856
Chromosome: 3R
Orientation: +
TSS: 27539771 Pattern: Embryo_supported
TSS: 27550733 Pattern: Embryo_supported
TSS: 27551504 Pattern: Embryo_specific

Gene: CG2671
Chromosome: 2L
Orientation: -
TSS: 18534 Pattern: Embryo_supported
TSS: 18567 Pattern: Embryo_specific
TSS: 21372 Pattern: Embryo_supported

Gene: CG31243
Chromosome: 3R
Orientation: +
TSS: 13757595 Pattern: Embryo_specific
TSS: 13769829 Pattern: Embryo_supported

Gene: CG3725
Chromosome: 2R
Orientation: -
TSS: 19437813 Pattern: Mixed
TSS: 19441060 Pattern: Embryo_supported,Schneider_cells_supported
TSS: 19441112 Pattern: Head_specific
TSS: 19441337 Pattern: Embryo_supported,Head_supported
TSS: 19441358 Pattern: Embryo_specific
TSS: 19443279 Pattern: Mixed

Gene: CG4898
Chromosome: 3R
Orientation: +
TSS: 11107274 Pattern: Mixed
TSS: 11116668 Pattern: Embryo_supported

Gene: CG8989
Chromosome: X
Orientation: -
TSS: 9000223 Pattern: Embryo_supported,Head_supported
TSS: 9001886 Pattern: Embryo_supported,Head_supported

Gene: CG9075
Chromosome: 2L
Orientation: +
TSS: 5981787 Pattern: Mixed
TSS: 5982600 Pattern: Head_supported

Gene: CG9261
Chromosome: 2L
Orientation: -
TSS: 6794595 Pattern: Embryo_supported,Head_supported
TSS: 6796779 Pattern: Embryo_supported,Head_supported

Gene: CG9553
Chromosome: 2L
Orientation: -
TSS: 5980098 Pattern: Embryo_supported,Head_supported
TSS: 5980982 Pattern: Embryo_supported,Head_supported

Appendix D Embryonic Expression Measured By Tiling Arrays

In Chapter 4, differences in Affymetrix tiling array fluorescence were used to determine TSS utilization at 12 time points for 2 hour increments during the first 24 hours of *Drosophila* embryogenesis. The gene, chromosome, orientation, and temporal pattern of utilization at each time point for the promoters of peaked TSSs and broad TSS cluster groups for the set of ten genes are listed.

Gene: CG_ID
Chromosome: #
Orientation: +/-
TSS or TSS_Cluster_Group: Location Embryonic_Utilization: Dev Stage 1,2,3,4,5,6,7,8,9,10,11,12 (T=Transcribed N = Not Transcribed)

Gene: CG10334
Chromosome: 2L
Orientation: -
TSS: 19568536 Embryonic_Utilization: T,T,T,T,T,T,T,T,T,T,T,T
TSS: 19571516 Embryonic_Utilization: N,N,N,N,N,N,N,N,N,N,N,N
TSS_Cluster_Group: 19573159_19573222 Embryonic_Utilization: N,T,T,T,T,T,T,N,N,N,N,N

Gene: CG1856
Chromosome: 3R
Orientation: +
TSS: 27539771 Embryonic_Utilization: N,N,T,N,T,T,N,N,N,N,T,N
TSS: 27550733 Embryonic_Utilization: T,T,T,T,T,T,T,T,T,N,N,N
TSS: 27551504 Embryonic_Utilization: N,N,T,T,T,T,N,N,N,N,N,N

Gene: CG2671
Chromosome: 2L
Orientation: -
TSS_Cluster_Group: 18534_18567 Embryonic_Utilization:
T,T,T,T,T,T,T,T,T,T,T,T
TSS: 21372 Embryonic_Utilization: T,T,T,T,T,T,T,T,T,T,T,T

Gene: CG31243
Chromosome: 3R
Orientation: +
TSS: 13757595 Embryonic_Utilization: N,N,N,N,N,N,N,T,T,T,N,T
TSS: 13769829 Embryonic_Utilization: N,N,N,T,T,T,T,T,T,N,N,N

Gene: CG3725
Chromosome: 2R
Orientation: -
TSS: 19437813 Embryonic_Utilization: N,N,N,N,N,N,N,N,N,N,N,N
TSS_Cluster_Group: 19441060_19441112 Embryonic_Utilization:
T,T,N,T,T,T,T,T,T,T,T
TSS_Cluster_Group: 19441337_19441358 Embryonic_Utilization:
N,N,N,N,N,N,N,N,N,N,N,N
TSS: 19443279 Embryonic_Utilization: N,N,N,N,N,N,N,T,T,T,T,T

Gene: CG4898
Chromosome: 3R
Orientation: +
TSS: 11107274 Embryonic_Utilization: N,N,N,N,N,N,N,T,N,T,N,N
TSS: 11116668 Embryonic_Utilization: N,T,N,N,N,N,N,N,N,N,N,N

Gene: CG8989
Chromosome: X
Orientation: -
TSS: 9000223 Embryonic_Utilization: N,N,N,N,N,N,T,T,N,N,N,N
TSS: 9001886 Embryonic_Utilization: T,T,T,T,T,T,T,T,T,T,T,T

Gene: CG9075
Chromosome: 2L
Orientation: +
TSS: 5981787 Embryonic_Utilization: T,T,T,T,T,T,T,T,T,T,T,T
TSS: 5982600 Embryonic_Utilization: T,T,T,T,T,T,N,T,N,N,N,N

Gene: CG9261
Chromosome: 2L
Orientation: -
TSS: 6794595 Embryonic_Utilization: T,N,N,T,T,T,N,N,N,N,N,N
TSS: 6796779 Embryonic_Utilization: N,N,N,N,N,N,N,T,T,T,T,T

Gene: CG9553
Chromosome: 2L
Orientation: -
TSS: 5980098 Embryonic_Utilization: T,T,T,T,T,T,T,T,T,T,T,T
TSS: 5980982 Embryonic_Utilization: T,T,T,T,T,T,T,N,N,N,N,N

Appendix E Most Highly Utilized Temporal Patterns During Embryogenesis

In Chapter 4, the temporal utilization of each promoter was evaluated using the tiling arrays for 12 time points, each time point corresponding to a 2 hour increment in development. Over all 12 time points, this created a pattern of transcribed (T) and non-transcribed (N) utilization for each promoter. The occurrence of the temporal patterns were cumulated across all promoters, regardless of their peaked or broad initiation pattern, and those that occurred at least 5 times are listed here. The patterns are grouped together according to the total number of periods of utilization. For instance, there are 12 patterns with 1 T, 17 patterns with 2 Ts, etc.

Pattern Frequency

N,N,N,N,N,N,N,N,N,N,N,N 1926

N,N,N,N,N,T,N,N,N,N,N,N 12
N,N,N,T,N,N,N,N,N,N,N,N 16
N,N,N,N,N,N,T,N,N,N,N,N 18
N,N,N,N,T,N,N,N,N,N,N,N 19
N,N,N,N,N,N,N,N,N,N,T,N 23
N,N,N,N,N,N,N,N,N,T,N,N,N 24
N,N,N,N,N,N,N,N,N,T,N,N 33
N,N,N,N,N,N,N,T,N,N,N,N 35
N,N,N,N,N,N,N,N,N,N,N,T 36
N,N,T,N,N,N,N,N,N,N,N,N 44
N,T,N,N,N,N,N,N,N,N,N,N 59
T,N,N,N,N,N,N,N,N,N,N,N 149

N,N,T,T,N,N,N,N,N,N,N,N 6
N,T,N,T,N,N,N,N,N,N,N,N 6
T,N,N,N,T,N,N,N,N,N,N,N 7
T,N,N,T,N,N,N,N,N,N,N,N 8
N,T,N,N,T,N,N,N,N,N,N,N 8
T,N,T,N,N,N,N,N,N,N,N,N 8
N,N,N,N,T,T,N,N,N,N,N,N 9
N,T,T,N,N,N,N,N,N,N,N,N 9
N,N,N,N,N,N,N,N,T,T,N,N 10
N,N,N,N,N,T,T,N,N,N,N,N 10
N,N,N,T,T,N,N,N,N,N,N,N 10
N,N,N,N,N,N,N,N,N,T,N,T 13
N,N,N,N,N,N,N,T,N,T,N,N 15
N,N,N,N,N,N,N,N,N,N,T,T 15
N,N,N,N,N,N,T,T,N,N,N,N 21
N,N,N,N,N,N,N,T,T,N,N,N 22
T,T,N,N,N,N,N,N,N,N,N,N 173

N,N,N,N,N,N,N,N,T,T,N,T 6
N,N,N,N,N,N,N,N,N,T,N,T,T 6
T,T,N,N,N,N,T,N,N,N,N,N 7
N,N,N,N,N,N,N,T,T,T,N,N 7
N,N,N,T,T,T,N,N,N,N,N,N 7
T,N,N,T,T,N,N,N,N,N,N,N 8
N,N,N,N,N,N,N,T,N,T,N,T 9
N,N,N,N,N,T,T,T,N,N,N,N 9
T,T,N,N,T,N,N,N,N,N,N,N 12
N,N,N,N,N,N,N,N,N,T,T,T 15
T,T,T,N,N,N,N,N,N,N,N,N 18
N,N,N,N,N,N,N,N,T,T,T,N 19
T,T,N,T,N,N,N,N,N,N,N,N 19

N,N,N,N,N,T,T,T,T,N,N,N 6
N,N,T,T,T,T,N,N,N,N,N,N 7
N,N,N,T,T,T,T,N,N,N,N,N 8
T,T,N,T,T,N,N,N,N,N,N,N 8
N,N,N,N,T,T,T,T,N,N,N,N 11

N,N,N,N,N,N,N,T,T,T,T,N 12
T,T,N,N,T,T,N,N,N,N,N,N 12
T,T,T,T,N,N,N,N,N,N,N,N 16
N,N,N,N,N,N,N,N,T,T,T,T 58

T,N,N,N,N,N,N,N,T,T,T,T 6
N,N,N,N,T,T,T,T,T,N,N,N 8
N,T,T,T,T,T,N,N,N,N,N,N 9
T,T,N,N,T,T,T,N,N,N,N,N 10
N,N,N,N,N,N,N,T,T,T,T,T 22
T,T,N,T,T,T,N,N,N,N,N,N 26
T,T,T,T,T,N,N,N,N,N,N,N 27

N,T,T,T,T,T,T,N,N,N,N,N 6
T,T,N,T,T,T,T,N,N,N,N,N 18
N,N,N,N,N,N,T,T,T,T,T,T 22
T,T,T,T,T,T,N,N,N,N,N,N 49

T,T,N,T,T,T,T,T,N,N,N,N 9
T,N,N,N,N,N,T,T,T,T,T,T 10
N,N,N,N,N,T,T,T,T,T,T,T 15
T,T,T,T,T,T,T,N,N,N,N,N 51

T,T,N,T,T,T,T,T,T,N,N,N 6
T,T,N,N,N,N,T,T,T,T,T,T 8
T,T,T,T,T,T,T,N,N,N,T,N 11
T,T,T,T,T,T,T,N,T,N,N,N 11
N,N,N,N,T,T,T,T,T,T,T,T 14
T,T,T,T,T,T,T,T,N,N,N,N 34

T,T,T,T,T,T,T,T,N,T,N,N 6
T,T,N,N,N,T,T,T,T,T,T,T 11
N,N,N,T,T,T,T,T,T,T,T,T 15
T,T,T,T,T,T,T,T,T,N,N,N 25

T,T,N,N,T,T,T,T,T,T,T,T 9
T,N,N,T,T,T,T,T,T,T,T,T 9
T,T,T,T,T,T,T,T,T,N,N,T 10
T,T,T,T,T,T,T,T,T,T,N,N 11

T,T,T,T,T,T,T,T,N,T,T,T 9
T,N,T,T,T,T,T,T,T,T,T,T 10
N,T,T,T,T,T,T,T,T,T,T,T 11
T,T,T,T,T,T,T,T,T,T,T,N 12
T,T,T,T,T,T,T,T,T,T,N,T 15
T,T,T,T,T,T,T,T,T,N,T,T 30
T,T,N,T,T,T,T,T,T,T,T,T 35

T,T,T,T,T,T,T,T,T,T,T,T 272

Appendix F Condition Specific Motifs

In Chapter 5, motifs were identified using the FREE and MEME search algorithms in the promoters with head, embryo, and testes associations. Column 5 notes the parameters (amplitude, mean, and standard deviation) returned by FREE for each of the Gaussians used to model the motif's position overrepresentation. The last column includes motif matches (and orientation) to the + strand of the motifs found by MEME (Library_MOTIF#), and the other known motif sets, including Ohler (Matrix #), FitzGerald (DM), 12 genomes (ME #), Tiffin (TIF#), Pause Button (Pause_Button), JASPAR (MA#), FlyReg (FlyReg #), and Fly (Fly #).

Search	Motif	Library	P-val	Gaussian Parameters	Motif Matches
FREE:	AGTCAG	Embryo	3.40E-141	(716.05 -4.46 0.54) (205.99 -0.52 0.56) (-21.57 -4.19 1.59)	Embryo_MOTIF16(+) FlyReg_tll
FREE:	CAGTGA	Embryo	6.60E-139	(444.92 -1.65 <.5) (69.25 19.01 <.5) (69.23 20.93 <.5) (69.22 0.97 <.5) (28.24 -2.00 <.5)	Fly_eyg TIFDMEM0000050
FREE:	CGCGCT	Embryo	2.30E-52	(275.52 14.00 <.5) (224.77 -14.82 <.5) (142.67 -38.00 0.79)	Fly_brk
FREE:	ATCATT	Embryo	1.30E-51	(445.23 -3.36 0.50) (316.62 -0.28 0.53)	none
FREE:	GGTTCA	Embryo Head	2.30E-51	(492.21 26.36 <.5) (212.58 -5.48 <.5) (68.34 -57.00 <.5) (68.34 -37.00 <.5)	Fly_Hr46 MAtrix9(+) ME139(-)
FREE:	CGTCAG	Embryo	1.90E-46	(550.28 -4.44 0.53) (135.12 -48.00 <.5)	none
FREE:	AGTTAA	Embryo	6.70E-45	(439.12 -0.69 <.5) (157.17 21.29 0.53)	Fly_eve TIFDMEM0000003
FREE:	GATTCA	Embryo	3.00E-	(493.62 -5.67 <.5)	none

			43	(136.82 -12.00 <.5)	
FREE:	CTCATT	Embryo	4.10E-43	(605.39 -3.82 <.5) (133.21 -52.00 <.5) (133.21 0.00 <.5)	FlyReg_kni
FREE:	CGCAAC	Embryo	6.90E-41	(342.90 26.00 <.5) (156.89 -55.31 0.58) (135.43 21.00 <.5)	MAtrix10(+)
FREE:	AGTTTC	Embryo	3.70E-39	(888.33 -0.32 0.64) (395.57 -6.55 0.53)	FlyReg_Hsf TIFDMEM0000008
FREE:	AGTTAT	Embryo	6.40E-36	(330.88 -0.38 0.60) (157.41 -4.26 0.53)	ME119(+) TIFDMEM0000047
FREE:	CTGCAGT	Embryo	2.90E-32	(616.23 -3.00 <.5) (132.14 -12.00 <.5)	Embryo_MOTIF17(+) ME117(-) TIFDMEM0000076
FREE:	CCCTGG	Embryo	3.20E-28	(26.82 -25.00 <.5) (19.83 14.00 <.5) (12.98 -22.00 <.5)	none
FREE:	CGGAGC	Embryo	4.60E-28	(270.14 26.00 <.5) (131.83 15.00 <.5)	none
FREE:	TTTGTC	Embryo	1.10E-27	(306.79 1.27 0.54) (132.52 -32.00 <.5) (132.52 -4.00 <.5)	Fly_br-Z1 MA0010(-) Testes_MOTIF3(+)
FREE:	CTTCGC	Embryo	7.60E-27	(268.70 -34.00 <.5) (130.39 -24.00 <.5) (130.39 4.00 <.5)	ME142(-)
FREE:	CAGTGTT	Embryo	1.60E-25	(630.70 -1.43 <.5) (263.28 -26.00 <.5)	Embryo_MOTIF2(-) Embryo_MOTIF6(-) MAtrix1(-) ME131(-) TIFDMEM0000109 TIFDMEM0000116
FREE:	GAGTGC	Embryo	1.70E-25	(280.16 20.98 <.5)	TIFDMEM0000007
FREE:	ATAAGC	Embryo	1.70E-25	(280.16 -27.00 <.5)	ME139(+) ME74(+)
FREE:	TTATTC	Embryo	8.10E-21	(291.25 -2.16 <.5)	FlyReg_ap Fly_hb MA0049(-) TIFDMEM0000002
FREE:	ATTGTG	Embryo	1.50E-18	(265.82 21.00 <.5) (196.66 18.00 <.5)	Fly_dsx
FREE:	CGCAGT	Embryo	3.40E-18	(479.21 -3.77 <.5) (210.03 -17.20 <.5)	FlyReg_Adf1
FREE:	ACGAAA	Embryo	1.00E-17	(268.70 31.00 <.5)	FlyReg_Deaf1 TIFDMEM0000077
FREE:	CTGCCA	Embryo	3.20E-17	(290.34 -50.86 <.5) (213.47 -19.78 <.5) (197.79 -10.00 <.5)	Embryo_MOTIF6(-) MAtrix8(-) TIFDMEM0000057
FREE:	GCCACA	Embryo	8.30E-16	(342.74 -23.92 <.5)	MAtrix8(-)
FREE:	CGGCAG	Embryo	1.20E-14	(307.67 -4.83 <.5) (148.88 20.69 0.51)	Embryo_MOTIF17(+) FlyReg_Mad Head_MOTIF15(+)

					MAtrix8(+)
FREE:	TGGCAT	Embryo	4.30E-14	(266.54 -4.00 <.5)	ME142(-)
FREE:	GTTGAA	Embryo	2.70E-13	(392.13 0.97 <.5)	Testes_MOTIF3(-)
FREE:	AAAATT	Embryo	1.40E-11	(3368.15 33.75 0.96) (3228.36 -17.00 0.51)	FlyReg_zen Fly_tll ME117(+) TIFDMEM0000089 Testes_MOTIF3(+)
FREE:	AAAAGC	Embryo	6.40E-13	(205.82 -28.85 1.50) (260.18 13.00 <.5)	MAtrix3(+) ME117(+) ME139(+) TIFDMEM0000051
FREE:	GAAAAG	Embryo	2.40E-26	(505.51 22.21 0.54) (468.69 35.00 <.5)	ME142(+) TIFDMEM0000018 Testes_MOTIF3(-)
FREE:	AMAGTCGC	Embryo	2.50E-45	(497.09 -2.53 <.5) (143.09 -18.10 <.5) (143.09 24.85 <.5)	FlyReg_Adf1 FlyReg_tll ME117(+) ME134(+) ME139(+) TIFDMEM0000101
FREE:	GACAGT	Embryo	1.10E-33	(425.77 -3.92 <.5)	none
FREE:	AGTCGA	Embryo	1.50E-27	(535.27 -0.28 0.60) (132.86 -43.00 0.77)	none
FREE:	AGTTAGTA	Embryo	3.40E-51	(619.43 -2.50 <.5) (136.78 0.98 <.5)	FlyReg_br-Z4 MAtrix7(-) ME119(-) ME131(-) ME142(-) TIFDMEM0000111
FREE:	AGTTGGCTC	Embryo	1.90E-37	(362.78 1.68 <.5) (183.62 -32.39 <.5) (136.77 -22.00 <.5) (136.77 29.00 <.5)	Embryo_MOTIF16(+) FlyReg_shn MAtrix6(+) ME117(-) TIFDMEM0000057 TIFDMEM0000058 TIFDMEM0000099
FREE:	AGCCAG	Embryo	7.70E-27	(268.70 -4.00 <.5) (130.39 -57.00 <.5) (130.39 10.00 <.5)	none
FREE:	TCAATTG	Embryo	1.60E-43	(611.85 -2.78 0.50) (133.22 20.00 <.5) (133.22 32.00 <.5)	FlyReg_ftz Fly_eve ME7(+)
FREE:	AACGGTYSTG	Embryo	2.90E-62	(608.22 24.31 <.5) (68.88 -46.00 <.5)	Embryo_MOTIF2(+) FlyReg_Aef1 FlyReg_ovo FlyReg_srp Fly_Kr Fly_prd-PD MAtrix9(+)

					ME117(-) ME139(-) TIFDMEM0000067
FREE:	AGTTCGWA	Embryo	2.00E-37	(361.29 0.44 0.55) (227.39 5.21 <.5) (134.76 -47.00 <.5)	FlyReg_Deaf1 ME131(-) ME142(+) Pause_Button(-) TIFDMEM0000036 TIFDMEM0000059
FREE:	GCGAACGG	Embryo	2.30E-37	(334.01 16.67 0.50) (349.79 21.39 0.69) (133.88 31.00 <.5)	FlyReg_ovo Fly_Kr MAtrix10(+) MAtrix9(-) Pause_Button(+)
FREE:	CGSTCGAA	Embryo	2.50E-34	(370.26 25.37 <.5) (211.64 15.77 0.77) (202.28 30.00 <.5)	FlyReg_Adf1 Head_MOTIF4(+) MAtrix9(-) Pause_Button(+) TIFDMEM0000059
FREE:	CACATCAG	Embryo	2.40E-34	(432.06 -4.36 0.51) (237.65 -1.31 0.58)	MAtrix5(+) ME117(+)
FREE:	AGCTGC	Embryo	2.40E-17	(286.50 -34.14 <.5)	Embryo_MOTIF17(+) Embryo_MOTIF6(-) Head_MOTIF15(-) ME134(+) ME35(+) TIFDMEM0000015
FREE:	TGCGWTCRGTT	Embryo Head	1.20E-48	(541.58 22.61 <.5) (140.27 16.97 <.5)	Embryo_MOTIF16(+) FlyReg_Deaf1 FlyReg_srp Head_MOTIF17(+) MAtrix5(+) MAtrix9(+) ME139(-) ME142(-) Pause_Button(+) TIFDMEM0000033 TIFDMEM0000071 TIFDMEM0000102
FREE:	ACGTGCGGTT	Embryo	1.10E-68	(847.79 24.48 <.5) (206.78 -36.99 <.5) (137.53 20.00 <.5)	FlyReg_HLHm5 Fly_gcm MAtrix9(+) ME117(-) ME124(-) ME139(-) ME39(+)
FREE:	YGGACGTG	Embryo	5.40E-99	(7958.30 26.66 0.52) (66.56 22.47 2.04)	DMp5(+) Embryo_MOTIF2(-) Fly_pros Head_MOTIF15(+) MAtrix9(-) ME124(-) ME134(+) ME139(+) Pause_Button(+)

					Embryo_MOTIF17(+) FlyReg_ovo Fly_Kr Head_MOTIF15(+) MAtrix10(+) ME117(+) Pause_Button(+)
FREE:	AARCGGAC	Embryo	4.20E-102	(435.71 24.65 0.53) (137.73 32.00 <.5) (68.57 14.00 <.5)	TIFDMEM0000071
					FlyReg_HLHm5 Fly_h MAtrix9(+)
FREE:	CGCGTGC	Embryo	4.20E-30	(288.99 19.15 <.5) (133.19 15.00 <.5) (133.19 24.00 <.5)	ME139(-)
FREE:	CGYTTC	Embryo	1.50E-24	(335.70 -7.00 <.5) (128.23 -41.00 0.53) (128.23 -16.00 <.5)	ME139(-) TIFDMEM0000071
					FlyReg_Abd-B FlyReg_Cf2-II FlyReg_hb Fly_Cf2 Fly_bin Fly_cad Fly_croc Fly_hb Head_MOTIF17(-) MA0013(+) MA0015(+) MA0015(-) MA0049(+) MAtrix3(+) MAtrix6(-) ME100(+) ME100(-) ME104(+) ME104(-) ME106(+) ME117(+) ME119(+) ME139(+) ME139(-) ME142(+) ME142(-) ME3(+) ME5(+) ME5(-) TIFDMEM0000034 TIFDMEM0000072 TIFDMEM0000083 TIFDMEM0000095 TIFDMEM0000104 TIFDMEM0000107 TIFDMEM0000111
FREE:	GGTATAAAWAC	Embryo Head	3.50E-49	(1295.97 -31.81 1.37)	Testes_MOTIF3(+) Testes_MOTIF3(-)
FREE:	CAGTAT	Embryo	8.70E-	(632.11 -1.48 <.5)	none

			57	(136.07 -4.00 <.5)	
FREE:	AGTATT	Embryo	1.10E-24	(305.39 -0.27 0.54) (131.78 -36.00 <.5) (131.78 -20.00 <.5)	TIFDMEM0000111
FREE:	TTCTCAG	Embryo	2.50E-55	(441.65 -4.68 <.5) (90.61 -8.50 0.60)	FlyReg_SuH Fly_Su_H_ ME142(-)
FREE:	TTCTTCT	Embryo	3.10E-13	(334.26 -45.00 <.5)	Embryo_MOTIF15(-) FlyReg_Hsf
FREE:	AGTCTTCA	Embryo	1.30E-61	(683.19 -0.30 <.5) (69.50 -15.04 <.5) (69.50 -7.02 <.5)	Fly_prd-PD Testes_MOTIF3(-)
FREE:	AGAAGA	Embryo	5.00E-11	(332.09 18.00 <.5)	Embryo_MOTIF15(+) FlyReg_Hsf
FREE:	TMWNWCWKTTTGAT	Embryo	7.00E-67	(743.52 -1.47 0.62) (218.69 -5.22 0.68)	DMp3(+) Embryo_MOTIF16(+) Embryo_MOTIF6(+) Embryo_MOTIF6(-) FlyReg_Abd-B FlyReg_Adf1 FlyReg_Aef1 FlyReg_BEAF-32 FlyReg_Deaf1 FlyReg_Dfd FlyReg_SuH FlyReg_ara FlyReg_bin FlyReg_br-Z3 FlyReg_br-Z4 FlyReg_ems FlyReg_gl FlyReg_gsb-n FlyReg_kni FlyReg_ovo FlyReg_p120 FlyReg_pan FlyReg_sd FlyReg_tll FlyReg_vnd FlyReg_z Fly_Aef1 Fly_Dref Fly_bin Fly_br-Z3 Fly_croc Fly_ems Fly_eve Fly_eyg Fly_grh Fly_ovo Fly_pan Fly_pho Fly_shn-ZFP1 Fly_tll Head_MOTIF11(-)

					Head_MOTIF4(+)
					Head_MOTIF7(+)
					MA0012(-)
					MA0013(-)
					MAtrix10(-)
					MAtrix4(+)
					MAtrix5(+)
					MAtrix7(-)
					MAtrix8(+)
					MAtrix8(-)
					ME100(+)
					ME104(+)
					ME106(-)
					ME117(+)
					ME117(-)
					ME119(+)
					ME119(-)
					ME121(-)
					ME131(-)
					ME134(+)
					ME139(+)
					ME139(-)
					ME142(+)
					ME142(-)
					ME3(+)
					ME3(-)
					ME4(-)
					ME63(+)
					ME7(+)
					ME83(+)
					ME83(-)
					TIFDMEM0000004
					TIFDMEM0000007
					TIFDMEM0000008
					TIFDMEM0000010
					TIFDMEM0000023
					TIFDMEM0000026
					TIFDMEM0000027
					TIFDMEM0000029
					TIFDMEM0000032
					TIFDMEM0000036
					TIFDMEM0000041
					TIFDMEM0000043
					TIFDMEM0000050
					TIFDMEM0000054
					TIFDMEM0000057
					TIFDMEM0000058
					TIFDMEM0000059
					TIFDMEM0000060
					TIFDMEM0000074
					TIFDMEM0000075
					TIFDMEM0000077
					TIFDMEM0000078
					TIFDMEM0000081
					TIFDMEM0000084

					TIFDMEM0000085 TIFDMEM0000088 TIFDMEM0000090 TIFDMEM0000093 TIFDMEM0000096 TIFDMEM0000099 TIFDMEM0000101 TIFDMEM0000106 TIFDMEM0000113 Testes_MOTIF3(+) Testes_MOTIF3(-)
FREE:	TCGACGTC	Embryo	2.60E-121	(384.29 27.69 <.5) (283.88 -57.67 <.5) (109.41 -51.50 <.5) (69.16 12.00 <.5)	Embryo_MOTIF11(+) ME16(+)
FREE:	CAKTTGMTT	Embryo Head	1.10E-59	(552.53 2.00 <.5) (149.89 -37.87 <.5) (142.06 9.00 0.79) (68.44 -58.00 <.5)	Embryo_MOTIF16(+) FlyReg_ara FlyReg_gsb-n MAtrix4(+) ME117(-) ME134(+) ME83(+) TIFDMEM0000036 TIFDMEM0000060 TIFDMEM0000093 TIFDMEM0000120
FREE:	GACGCAT	Embryo	1.00E-47	(275.90 25.00 <.5) (206.85 29.00 <.5) (69.87 -32.95 <.5) (69.87 4.94 <.5)	Fly_gcm ME134(+)
FREE:	AGCAAC	Embryo	6.10E-30	(313.48 -23.28 0.54) (255.02 25.64 0.74) (187.81 7.51 0.55) (133.16 16.00 <.5)	none
FREE:	GCGTCG	Embryo	4.60E-28	(361.58 12.12 0.75) (220.58 16.80 <.5) (131.65 23.00 <.5)	Embryo_MOTIF11(+) FlyReg_Mad MAtrix9(-)
FREE:	CAAATC	Embryo	2.50E-13	(294.76 -6.26 0.54) (195.76 -12.00 <.5)	ME117(+)
FREE:	TTGATT	Embryo	1.50E-12	(332.09 2.00 <.5)	FlyReg_ara Fly_prd-HD ME117(-) ME63(-) TIFDMEM0000060
FREE:	AAATGC	Embryo	2.60E-12	(280.61 -28.87 <.5)	ME117(+) ME139(+) TIFDMEM0000090
FREE:	CATTCAG	Head Embryo	7.00E-126	(107631.87 -4.25 <.5) (57647.29 0.32 <.5) (7.89 0.04 1.66)	Embryo_MOTIF16(+) FlyReg_p120
FREE:	GGAATTGT	Head Embryo	1.00E-90	(86.78 5.50 0.59) (29.62 -20.00 <.5)	FlyReg_Dfd FlyReg_Eip74EF

				(14.76 -15.00 <.5)	FlyReg_ftz Fly_dl-B TIFDMEM0000048
FREE:	TTCGGT	Head	2.50E-87	(58.96 -43.00 <.5) (44.39 23.99 <.5) (29.24 -60.00 <.5) (14.38 -9.00 <.5)	FlyReg_Deaf1 Head_MOTIF17(+) MAtrix9(+) TIFDMEM0000071
FREE:	AACATT	Head	1.10E-55	(74.15 -3.00 <.5) (48.07 -8.20 0.75) (14.71 5.00 <.5)	FlyReg_p120 TIFDMEM0000064 Testes_MOTIF3(+)
FREE:	GTTGAC	Head	1.20E-52	(68.27 29.46 0.63)	Testes_MOTIF3(-)
FREE:	TTGTCT	Head	5.00E-50	(85.76 7.43 <.5) (29.59 -6.00 <.5) (29.59 29.00 <.5) (29.59 33.00 <.5)	Fly_br-Z1 MA0010(-) ME117(-) Testes_MOTIF3(+)
FREE:	GCATTT	Head Embryo	3.20E-48	(86.04 -17.39 <.5) (55.65 23.68 <.5) (29.56 29.00 <.5) (14.70 -6.00 <.5) (14.70 11.00 <.5)	Fly_gcm ME117(-) ME139(-) TIFDMEM0000090
FREE:	AAGCAG	Head	1.90E-42	(83.88 -27.55 <.5) (30.18 -17.00 <.5)	none
FREE:	TGGATT	Head	8.40E-42	(67.45 -41.54 0.63)	ME117(-)
FREE:	CTTTGGA	Head	8.40E-42	(67.45 -43.54 0.63)	ME142(-)
FREE:	TTCACT	Head	5.60E-39	(84.99 -33.55 <.5) (44.19 -2.96 <.5)	Fly_eyg TIFDMEM0000050
FREE:	CTTTCG	Head	2.80E-37	(59.09 -49.00 <.5) (29.39 -37.00 <.5)	ME142(-)
FREE:	CTATTT	Head	1.20E-34	(96.60 15.24 0.56)	FlyReg_br-Z1 Fly_br-Z2 MA0011(+) ME119(+) ME142(+) TIFDMEM0000022
FREE:	AGTTTGG	Head	6.70E-32	(71.52 -2.55 0.58)	Head_MOTIF4(+) ME117(-)
FREE:	CAGCAC	Head	4.40E-28	(58.04 -3.00 <.5) (28.32 10.00 <.5)	Embryo_MOTIF6(+) MAtrix7(+) TIFDMEM0000109
FREE:	TTTCTT	Head	2.40E-26	(63.42 -56.17 <.5)	Embryo_MOTIF15(-) MA0013(-) TIFDMEM0000011
FREE:	TTTCAG	Head	2.20E-24	(57.89 -4.00 <.5) (28.17 -55.00 <.5)	TIFDMEM0000027
FREE:	TTTTGA	Head	6.70E-23	(65.92 -13.27 0.54) (28.47 -57.00 <.5)	FlyReg_pan ME142(-) TIFDMEM0000113 Testes_MOTIF3(+)
FREE:	CTGGTC	Head	1.20E-21	(57.73 -6.00 <.5) (28.02 -30.00 <.5)	Embryo_MOTIF2(+)
FREE:	TCGCTT	Head	9.90E-	(57.73 2.00 <.5)	none

			18		
FREE:	TTTTGC	Head	9.90E-18	(57.73 28.00 <.5)	MA0049(-) TIFDMEM0000089
FREE:	AAATTC	Head	1.40E-25	(66.79 2.72 0.54)	Fly_tll Head_MOTIF11(+) ME117(+) ME139(+) TIFDMEM0000078
FREE:	AATTCA	Head	1.20E-52	(68.27 -35.54 0.63)	Fly_tll ME117(+) TIFDMEM0000078
FREE:	ATTAATT	Head	2.10E-38	(79.74 -37.73 0.61)	FlyReg_Dfd FlyReg_ara Fly_tll MA0013(-) ME100(+) ME119(+) ME7(+) TIFDMEM0000004 TIFDMEM0000113
FREE:	CCAGTCGS	Head	4.30E-89	(84.21 11.75 0.50) (14.57 -60.00 <.5) (14.57 0.00 <.5)	FlyReg_Adf1 ME134(+) ME139(+) TIFDMEM0000101
FREE:	CNTCAGTTC	Head	3.30E-97	(73.82 -3.00 <.5) (44.10 1.00 <.5)	Embryo_MOTIF16(+) FlyReg_br-Z4 Fly_eyg MAtrix4(+) MAtrix5(+) ME131(-) TIFDMEM0000007 TIFDMEM0000036 TIFDMEM0000060
FREE:	TCGGCAA	Head Embryo	4.80E-157	(953.65 14.66 <.5) (148.60 -18.00 <.5) (54.90 15.00 <.5)	FlyReg_Deaf1 MAtrix8(+) ME117(+)
FREE:	TTCATTCGC	Head Embryo	3.60E-61	(192.69 -2.00 <.5)	DMp3(+) Embryo_MOTIF16(+) FlyReg_ara Fly_croc Fly_tll MAtrix4(+) ME104(+) ME139(+) ME142(-) TIFDMEM0000077 TIFDMEM0000102 TIFDMEM0000106
FREE:	CAACAA	Head Embryo	1.30E-115	(59.44 16.00 <.5) (42.58 -27.38 <.5) (23.80 -51.50 <.5) (14.86 13.00 <.5) (14.86 24.00 <.5)	FlyReg_Aef1 Fly_Aef1 MA0010(+) ME117(+) ME142(+) ME83(+)

					Testes_MOTIF3(-) TIFDMEM0000065
FREE:	TTTGTTT	Head	2.00E-37	(74.49 -44.33 <.5) (50.59 -48.60 0.73) (43.65 15.00 <.5)	FlyReg_bin Fly_Aef1 Head_MOTIF7(+) MA0010(-) MA0013(-) ME117(-) TIFDMEM0000084 Testes_MOTIF3(+)
FREE:	TTTTCG	Head	1.10E-13	(64.68 5.74 0.54)	TIFDMEM0000012
FREE:	GTTTTC	Head	1.30E-61	(106.67 5.11 <.5) (28.79 -23.00 <.5) (28.79 35.00 <.5) (13.93 8.00 <.5) (13.93 32.00 <.5)	FlyReg_dl MA0023(+) TIFDMEM0000018
FREE:	ATTTTTTAT	Head	1.00E-25	(67.01 20.45 0.61)	FlyReg_Abd-B FlyReg_hb Fly_croc Fly_hb Head_MOTIF17(+) MA0013(-) MA0049(-) MAtrix3(-) MAtrix6(+) ME104(+) ME106(-) ME117(-) ME119(-) ME139(-) ME142(-) ME5(-) TIFDMEM0000002 TIFDMEM0000049 TIFDMEM0000072 TIFDMEM0000107 Testes_MOTIF3(+)
FREE:	TTTATTTAT	Head	2.90E-22	(60.49 -31.80 <.5) (29.20 -29.24 0.94)	FlyReg_Abd-B FlyReg_ap Fly_bin Fly_cad MA0010(-) MA0013(-) MA0049(-) ME100(+) ME104(-) ME106(+) ME119(+) ME139(+) ME139(-) ME142(+) ME3(+) ME5(+) ME5(-)

					TIFDMEM0000002 TIFDMEM0000072 TIFDMEM0000104 Testes_MOTIF3(+)
FREE:	TTTTAGTC	Head	1.10E-54	(88.88 -3.00 <.5) (83.66 -50.52 0.58)	FlyReg_Abd-B FlyReg_br-Z4 Fly_croc MA0012(-) ME106(-) ME119(-) TIFDMEM0000029 TIFDMEM0000041 Testes_MOTIF3(+)
FREE:	NGGKCACACWGY	Head	2.90E-53	(78.75 -12.56 0.56) (29.60 -0.00 <.5) (29.60 26.00 <.5)	Embryo_MOTIF16(+) Embryo_MOTIF2(+) FlyReg_ftz-f1 Fly_Hr46 Fly_usp MA0016(+) MAtrix1(+) ME121(+) ME131(+) ME89(+) TIFDMEM0000116
FREE:	AGCTCGAGY	Head	3.50E-63	(160.64 -5.56 0.51)	Head_MOTIF4(+) Head_MOTIF4(-) TIFDMEM0000010
FREE:	GAGCTGCAA	Head	2.50E-123	(11047.37 -19.54 <.5) (4457.68 -5.00 <.5) (-60.83 -19.95 0.78) (653.28 -20.00 0.60)	Embryo_MOTIF17(+) Embryo_MOTIF6(-) FlyReg_slbo Head_MOTIF15(-) Head_MOTIF4(-) MAtrix7(-) ME117(-) ME134(+) ME35(+) TIFDMEM0000015 TIFDMEM0000076 TIFDMEM0000079
FREE:	ACAGCTC	Head Embryo	2.20E-43	(83.97 -10.58 <.5) (29.76 8.99 <.5)	Embryo_MOTIF6(+) Head_MOTIF4(+) MAtrix5(+) MAtrix7(+) ME35(+) TIFDMEM0000079
FREE:	GCTGTT	Head	1.10E-24	(65.62 -33.27 0.54) (28.31 -16.00 <.5) (28.31 12.00 <.5)	MAtrix5(+) ME35(-) ME117(-) Fly_ovo TIFDMEM0000079
MEME:	YGGTCACACTG	Embryo_MOTIF1 Head_MOTIF1			MAtrix1 TIFDMEM0000116
MEME:	CCAYCTCTAG	Embryo_MOTIF3			Fly_pros
MEME:	AACAGCTGWTTG	Embryo_MOTIF4			MAtrix5

		Head_MOTIF9			ME35 ME133 TIFDMEM0000079
MEME:	RNTATCGATARC	Embryo_MOTIF5			MAtrix2 FlyReg_BEAF-32 Fly_Dref TIFDMEM0000005
MEME:	KTTCAGTTNRNTTT	Embryo_MOTIF7			MAtrix4 ME120 TIFDMEM0000036
MEME:	GYGTGTGTGTGTGTG	Embryo_MOTIF9 Head_MOTIF16			ME89 ME121 ME131
MEME:	TTTGTTKTTGTTKTK	Embryo_MOTIF10			ME112 ME117 FlyReg_Aef1 Fly_Aef1 TIFDMEM0000065
MEME:	MAAWANMAAAAMAAW	Embryo_MOTIF13			MA0049 ME112 ME117 Fly_br-Z1 TIFDMEM0000094
MEME:	TGGTATTTTTCMR	Embryo_MOTIF14			FlyReg_dl Fly_dl-A TIFDMEM0000091
MEME:	ACAGAGYTGCCANMY	Embryo_MOTIF18			TIFDMEM0000042
MEME:	GCGTTTNCCAACACT	Embryo_MOTIF19			TIFDMEM0000109
MEME:	ACSTGCTTGCATTTT	Head_MOTIF2			MA0086 FlyReg_nub TIFDMEM0000048
MEME:	GTCTTYSKATTWWTT	Head_MOTIF3			MA0094 FlyReg_br-Z1 TIFDMEM0000070
MEME:	TCGGCMAGCCATCGT	Head_MOTIF8			none
MEME:	GCTATAWAAR	Head_MOTIF10			MAtrix3 ME5 Fly_croc TIFDMEM0000083
MEME:	AAATAYAAAWWW	Testes_MOTIF1			MA0011 FlyReg_br-Z1 Fly_br-Z2 TIFDMEM0000095

Appendix G Supplementary Methods

The experimental and computational techniques performed in Chapter 6 are described here.

Oligo(dT) selection

Dynabeads Oligo(dT)25 from Invitrogen was used to enrich poly(A)+ RNAs. Briefly, 150 μg total RNA was resuspended in 400 μl binding buffer (20 mM Tris-HCl, pH 7.5, 1.0 M LiCl, 2 mM EDTA, 1% LiDS, 0.1% Trion X-100) and heated at 65°C for 2 min to disrupt RNA secondary structures. After snap cool down, 200 μl Dynabeads was added followed by incubation at 50°C for 5 min. We found that incubation at a higher temperature helps remove the non-specific binding of ribosomal RNA. The resulting beads were then washed 3 times with Washing Buffer B (10 mM Tris-HCl, pH 7.5, 0.15 M LiCl, 1 mM EDTA, 0.1% LiDs, 0.1% Triton X-100). The RNA fraction bound to the beads was then eluted with 10 mM Tris-HCl (RNase-free) by heating at 75-80°C for 2 min. The entire poly(A) selection procedure was repeated one more time. The final RNA sample was further purified by MinElute kit (QIAGEN) to remove lithium salt, which otherwise would affect the activity of Bacterial Alkaline Phosphatase (BAP) in the subsequent step.

BAP/TAP treatment

1-2 μg of poly(A)+ RNA was first dephosphorylated in a 100 μl reaction (2.4 units Bacterial Alkaline Phosphatase (BAP; Takara), 50 mM Tris-HCl buffer (pH 9.0), 1 mM MgCl2, 50 mM NaCl and 100 units of RNasin Ribonuclease inhibitor (Promega)) at 37°C for 40 min. After phenol/chloroform extraction and ethanol precipitation, the resulting RNAs were treated with 20 units of TAP (Tobacco Acid Pyrophosphatase; Epicentre) in a 100 μl reaction (50 mM NaOAc (pH 6.0), 1 mM EDTA, 0.1% β-ME, 0.01% Triton X-100 and 100 unit RNasin) at 37°C for 1 hr. The reaction mixture was then extracted twice with phenol/chloroform and the RNA fragments were ethanol precipitated for downstream linker ligation.

Linker ligation

A chimeric linker (5′-CTC AAG CTT CTA ACG ATG TAC GCT CGrA rGrUrC rCrArA rC-3′) was ligated to poly(A)+ RNA with BAP/TAP treatment. The ligation was performed in 100 μl reaction including the recovered RNA, 60 pmol PAGE-purified linker (IDT), 200 units of T4 RNA ligase1 (NEB), 50 mM Tris-HCl (pH 7.8), 10 mM MgCl2, 1 mM ATP, 10 mM DTT, 25% PEG8000 and 100 units of RNasin Ribonuclease inhibitor (Promega). The reaction mixture was incubated overnight at room temperature, followed by phenol/chloroform extraction to remove both protein and

PEG8000. Ethanol precipitation was then performed to recover the RNA by adding 1/10 volume of NaOAc (pH 5.2) and 30 μg of GlycoBlue (Ambion).

Reverse Transcription

Random primer with a common sequence (5'-GCG GCT GAA GAC GGC CTA TCC GAC NNN NNN-3') was used to initiate the reverse transcription. Linker ligated RNAs were reversely transcribed in 40 μl reaction, which contains 20 pmol random primer, 2 nmol dNTP (Bioline), 240 ng actinomycin D, 80 units of RNasin Ribonuclease inhibitor (Promega), 400 units of Superscript III reverse transcriptase (Invitrogen), 50 mM Tris-HCl (pH 8.3), 75 mM KCl, 5 mM MgCl2 and 5 mM DTT. The reaction mixture was incubated at 25°C for 5min and 50°C for 1 hour, followed by heat inactivation at 70°C for 15min.

Circularization and rolling circle amplification

First-strand cDNAs were amplified by 2-5 cycles of PCR with high-fidelity DNA polymerase (Finnzymes) to generate blunt-end dsDNAs. Circularization of dsDNAs was achieved by bridge ligation. A 30 μl reaction was assembled which contains 1x Ampligase buffer (Epicentre), 7.5 units of OptiKinse (USB), 5 mM ATP and 1 mM DTT. The reaction was incubated at 37°C for 30 min, followed by 95 °C for 2 min to inactivate the kinase. 9 pmol "bridge" oligonucleotide (5'-GCC GTC TTC AGC CGC CTCA AGC

TTC TAA CGA TGT ACG-3') and 7.5 units of Ampligase (Epicentre) were then added. Ligation was performed with 5 cycles of 95°C for 30 sec, 68°C for 2 min, 55°C for 1 min and 60°C for 5 min, followed by 5 cycles of 95°C for 30 sec, 65°C for 2 min, 55°C for 1 min and 60°C for 5 min. To remove excess oligonucleotides and unligated DNA fragments, we added 3 μl Exonuclease I (NEB) and 0.6 μl Exonuclease III (NEB) to the ligation reaction and the mixture was incubated at 37°C for 45 min followed by 80°C for 20 min. This removes all linear DNA fragments (ssDNAs and dsDNAs) and the remaining circular DNAs were further amplified by rolling circle amplification (RCA). RCA was performed in four 20 μl reactions, each of which contains 2 μl DNA circles, 20 mM dNTP (Epicentre), 4 μg BSA (NEB), 200 pmol of random hexamer (5'-NNNN*N*N-3'; * = phosphothiol group), 10 units phi29 DNA polymerase (NEB), 2 μl DMSO and 1 x phi29 reaction buffer (NEB). The RCA reactions were incubated at 10°C for 10 min, 28°C for 16 hours and heat inactive at 65°C for 10 min.

Library quality control (QC)

Because the linker sequence has a built-in XhoI site, we use XhoI digestion to check the specificity of RCA amplification. Typically, the majority of the RCA products can be digested by XhoI and show an evenly-distributed smear centered around 300-400 bp. As another QC step, Sanger sequencing was also performed to check the quality before Illumina paired-end sequencing. After PCR amplification, a portion of the final

library was A-tailed and ligated into T-easy vector (Promega). We followed the standard procedure recommended by the manufacturer and the T7 sequence was used as sequencing primer. In general, ~80-90% of the clones mapped to known TSS or its surrounding regions, consistent with the data generated by Illumina/Solexa sequencing.

Trimming of sequence adaptor

The raw data we obtained are paired 35mer reads, each of which consists of a ~20nt tag (derived from MmeI digestion) followed by a 16nt linker sequence. The linker sequences were trimmed from the reads and subsequently used to identify which end of the transcript the read was from. Although both the 5' and 3' linker contain a MmeI site (5'-TCCAAC-3', 5'-TCCGAC-3', respectively), the sequences beyond their MmeI sites are completely different, thereby allowing for reliable determination of read directionality. Linker sequences were identified and trimmed such that there were no more than 2 mismatches/indels between the 3' end of the read and one of the two complete linker sequences. Read pairs in which either read from a pair failed to meet the linker sequence requirements were discarded from further analysis.

Correlation between read counts and microarray expression values

Microarray expression values were collected from the NCBI GEO repository (Barrett et al. 2009): dataset accession number GSE11880. The data originated from three arrays containing expression values for wild type *Drosophila melanogaster* mixed embryos of stages 0-11 (GEO accession numbers: GSM300072, GSM300074, and GSM3000). The mean value across all three replicates, after median background subtraction, was used

for our analysis. Genes with an average 'signal minus background' value less than 0 were given a log2-transformed expression value of 0. The total number of 3' reads from the aligned pairs that mapped to the transcribed region of each gene was used for comparison with the microarray data. The Pearson correlation coefficient was calculated across 10,101 genes that had at least one read-pair mapped to them and was present in the microarray data. This analysis was done twice, the first was performed on all read-pairs that mapped to a gene; the second analysis used only non-redundant read pairs.

Tag clustering strategy by F-Seq

F-Seq was used to perform the tag clustering (Boyle et al. F-seq: A feature density estimator for high-throughput sequence tags 2008). The 'fragment size' parameter refers to the size of the fragment that needs to be clustered and analyzed. It has been shown that the fragment size should be set to '1bp' (equivalent to a value of 0) for data sets where one end (in our case, the 5' end) of the sequence represents the point of enrichment. The 'feature length' parameter, on the other hand, controls the kernel density estimate bandwidth. The 'feature length' was set at 30, which means that the standard deviation of the Gaussian density estimate of a location has a value of 5bp

Core promoter motif position weight matrices used in the PATSER search are listed in Appendix B.

Primer pairs for validation of 5′ capped reads in CDS:

Gene name in FlyBase	Forward junction primer	Reverse GSP primer
FBgn0003870	GCTCGAGTCCAACCATACTTAAGGAT	ACCTCGGTGAGGCCCTTGATG
FBgn0014269	TCGAGTCCAACCCGAAAAGGAT	TCCTCGTCCACCTCATCGTTGA
FBgn0020443	CTCGAGTCCAACCACTTCTGTTGG	TGGCTGCATCGCAGAGAACAA
FBgn0024841	GCTCGAGTCCAACAGTTCTGTATTCG	TGCGCTTTTTCGCTGTGATTG
FBgn0026188	CTCGAGTCCAACCAGACACAGGA	GGCTTCTCGCGGATCTTGTCC
FBgn0029629	TCGAGTCCAACCGGAGTCGAT	TGCACGTAGGCGAATCCCTTG
FBgn0030341	CTCGAGTCCAACCATCAGTTCGTT	ATGTTCATCCACACCGGGCAG
FBgn0031769	GCTCGAGTCCAACCAAGTTGGC	CGTATACGGTTCCCGTGTCAGTGA
FBgn0035121	CTCGAGTCCAACCGATTTTGGAA	CGCGGCTTTGTACTGGTCCAC
FBgn0037301	CTCGAGTCCAACCATTCTTCGAG	GCTCGCCGTCAATGGAATTGA
FBgn0037707	GCTCGAGTCCAACCAATAGGTTCAA	TCGATCAGTGTCCTCTAACGAGAGC
FBgn0051729	CTCGAGTCCAACGAGCACTGG	CTGGCATCGATGCTATGTAGCTCC

Primer pairs for validation of novel TSSs

Gene name in FlyBase	Forward junction primer	Reverse GSP primer
FBgn0028537	CTCGAGTCCAACATTCACATTAGCA	CCGCCGAATCCAAACTGAGGT
FBgn0033113	CTCGAGTCCAACAGTCAATTCCG	TGACCCTATGGGTGCGCTGAT
FBgn0033068	CTCGAGTCCAACATTCGTTACCGT	GGCGTTATTTCCGACGTCTTTTGT
FBgn0033688	GCTCGAGTCCAACATTCTATAAATGG	TTTTCTGCCTTTCTTTGGCCG
FBgn0260442	CTCGAGTCCAACACTTTCGTGC	GATGTCGGTTCCATATGGCGG
FBgn0004228	CTCGAGTCCAACACAGTCGCTTT	CGTTGCACATGATGATAATCCTCG
FBgn0037410	CGAGTCCAACTTTTTTTTTTTTTGCT	CGCCTGGGGTGATTTTGGTTA
FBgn0085320	GCTCGAGTCCAACAGTTATCCATTT	GTACATTATCAGCAATTCCTTCTGTTCAA
FBgn0030136	CTCGAGTCCAACAGTTTGCTCATT	AGGAACTCGACCTTCACCTGGGT
FBgn0024366	GAGTCCAACAGTACGCTGCGAA	AGCGGACTCGGTTTTGTTTCGT

References

Adelman, K., M. A. Kennedy, S. Nechaev, D. A. Gilchrist, G. W. Muse, Y. Chinenov, and I. Rogatsky. 2009. Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. Proc Natl Acad Sci U S A 106, no. 43: 18207-12.

Adkins, N. L., T. A. Hagerman, and P. Georgel. 2006. Gaga protein: A multi-faceted transcription factor. Biochem Cell Biol 84, no. 4: 559-67.

Ahsan, B., T. L. Saito, S. Hashimoto, K. Muramatsu, M. Tsuda, A. Sasaki, K. Matsushima, T. Aigaki, and S. Morishita. 2009. Machibase: A drosophila melanogaster 5'-end mRNA transcription database. Nucleic Acids Res 37, no. Database issue: D49-53.

Aizer, A. and Y. Shav-Tal. 2008. Intracellular trafficking and dynamics of p bodies. Prion 2, no. 4: 131-4.

Akhtar, W. and G. J. Veenstra. 2009. Tbp2 is a substitute for tbp in xenopus oocyte transcription. BMC Biol 7: 45.

Albert, I., T. N. Mavrich, L. P. Tomsho, J. Qi, S. J. Zanton, S. C. Schuster, and B. F. Pugh. 2007. Translational and rotational settings of h2a.Z nucleosomes across the saccharomyces cerevisiae genome. Nature 446, no. 7135: 572-6.

Alberts, B. , A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. 2002. Molecular biology of the cell: Garland Science.

Alkema, W. B., O. Johansson, J. Lagergren, and W. W. Wasserman. 2004. Mscan: Identification of functional clusters of transcription factor binding sites. Nucleic Acids Res 32, no. Web Server issue: W195-8.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215, no. 3: 403-10.

Andrews, J., M. Smith, J. Merakovsky, M. Coulson, F. Hannan, and L. E. Kelly. 1996. The stoned locus of drosophila melanogaster produces a dicistronic transcript and encodes two distinct polypeptides. Genetics 143, no. 4: 1699-711.

Annunziato, A.T. 2008. DNA packaging: Nucleosomes and chromatin. Nature Education 1, no. 1.

Anwar, F., S. M. Baker, T. Jabid, M. Mehedi Hasan, M. Shoyaib, H. Khan, and R. Walshe. 2008. Pol II promoter prediction using characteristic 4-mer motifs: A machine learning approach. BMC Bioinformatics 9: 414.

Aoyagi, N. and D. A. Wassarman. 2001. Developmental and transcriptional consequences of mutations in drosophila taf(II)60. Mol Cell Biol 21, no. 20: 6808-19.

Araujo, S. J., C. Cela, and M. Llimargas. 2007. Tramtrack regulates different morphogenetic events during drosophila tracheal development. Development 134, no. 20: 3665-76.

Aravind, L. and D. Landsman. 1998. At-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Res 26, no. 19: 4413-21.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat Genet 25, no. 1: 25-9.

Bae, Y. J., K. S. Park, and S. J. Kang. 2003. Genomic organization and expression of parkin in drosophila melanogaster. Exp Mol Med 35, no. 5: 393-402.

Bailey, T. L. and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.

Bardin, AJ. , R. Le Borgne, and F. Schweisguth. 2004. Asymmetric localization and function of cell-fate determinants: A fly's view. Current Opinion Neurobiology 14, no. 1: 6-14.

Barrett, T., D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. 2009. Ncbi geo: Archive for high-throughput functional genomic data. Nucleic Acids Res 37, no. Database issue: D885-90.

Barski, A., S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. 2007. High-resolution profiling of histone methylations in the human genome. Cell 129, no. 4: 823-37.

Bartel, D. P. 2004. Micrornas: Genomics, biogenesis, mechanism, and function. Cell 116, no. 2: 281-97.

Beckett, D. 2001. Regulated assembly of transcription factors and control of transcription initiation. J Mol Biol 314, no. 3: 335-52.

Beller, M., D. Riedel, L. Jansch, G. Dieterich, J. Wehland, H. Jackle, and R. P. Kuhnlein. 2006. Characterization of the drosophila lipid droplet subproteome. Mol Cell Proteomics 5, no. 6: 1082-94.

Benoit, B., C. H. He, F. Zhang, S. M. Votruba, W. Tadros, J. T. Westwood, C. A. Smibert, H. D. Lipshitz, and W. E. Theurkauf. 2009. An essential role for the RNA-binding protein smaug during the drosophila maternal-to-zygotic transition. Development 136, no. 6: 923-32.

Berendzen, K. W., K. Stuber, K. Harter, and D. Wanke. 2006. Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional

disequilibrium in eukaryote genomes using frequency distribution curves. BMC Bioinformatics 7: 522.

Bergman, C. 2007. *Drosophila* curated transcription factor motifs. Accessed. Available from http://www.bioinf.manchester.ac.uk/bergman/data/motifs.

Bergman, C. M., J. W. Carlson, and S. E. Celniker. 2005. Drosophila dnase i footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, drosophila melanogaster. Bioinformatics 21, no. 8: 1747-9.

Bernstein, B. E., A. Meissner, and E. S. Lander. 2007. The mammalian epigenome. Cell 128, no. 4: 669-81.

Bertone, P., V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. 2004. Global identification of human transcribed sequences with genome tiling arrays. Science 306, no. 5705: 2242-6.

Biosciences, BD. 2003. BD SMART RACE cDNA amplification kit user manual, Palo Alto, CA.

Blanchette, M., A. R. Bataille, X. Chen, C. Poitras, J. Laganiere, C. Lefebvre, G. Deblois, V. Giguere, V. Ferretti, D. Bergeron, B. Coulombe, and F. Robert. 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 16, no. 5: 656-68.

Blumenthal, T. 2004. Operons in eukaryotes. Brief Funct Genomic Proteomic 3, no. 3: 199-211.

Boeger, H., D. A. Bushnell, R. Davis, J. Griesenbeck, Y. Lorch, J. S. Strattan, K. D. Westover, and R. D. Kornberg. 2005. Structural basis of eukaryotic gene transcription. FEBS Lett 579, no. 4: 899-903.

Boguski, M. S., T. M. Lowe, and C. M. Tolstoshev. 1993. dbEST--database for "Expressed sequence tags". Nat Genet 4, no. 4: 332-3.

Boyle, A. P., S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. 2008. High-resolution mapping and characterization of open chromatin across the genome. Cell 132, no. 2: 311-22.

Boyle, A. P., J. Guinney, G. E. Crawford, and T. S. Furey. 2008. F-seq: A feature density estimator for high-throughput sequence tags. Bioinformatics 24, no. 21: 2537-8.

Brogna, S. and M. Ashburner. 1997. The adh-related gene of drosophila melanogaster is expressed as a functional dicistronic messenger RNA: Multigenic transcription in higher organisms. Embo J 16, no. 8: 2023-31.

Brudno, M., C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou. 2003. Lagan and multi-lagan: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13, no. 4: 721-31.

Bryne, J. C., E. Valen, M. H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. 2008. JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. Nucleic Acids Res 36, no. Database issue: D102-6.

Bucher, P. and E. N. Trifonov. 1986. Compilation and analysis of eukaryotic pol II promoter sequences. Nucleic Acids Res 14, no. 24: 10009-26.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol 268, no. 1: 78-94.

Burges, CJC. 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2: 121-67.

Burke, T. W. and J. T. Kadonaga. 1996. Drosophila tfiid binds to a conserved downstream basal promoter element that is present in many tata-box-deficient promoters. Genes Dev 10, no. 6: 711-24.

Butler, J. E. and J. T. Kadonaga. 2001. Enhancer-promoter specificity mediated by dpe or tata core promoter motifs. Genes Dev 15, no. 19: 2515-9.

Camos, S and N Badia. 2006. *Drosophila* phylogeny. Accessed. Available from http://bioinformatica.upf.edu/2006/projectes06/3.7/Drosophila-phylogeny.jpg.

Carninci, P.,T. Kasukawa,S. Katayama,J. Gough,M. C. Frith,N. Maeda,R. Oyama,T. Ravasi,B. Lenhard,C. Wells,R. Kodzius,K. Shimokawa,V. B. Bajic,S. E. Brenner,S. Batalov,A. R. Forrest,M. Zavolan,M. J. Davis,L. G. Wilming,V. Aidinis,J. E. Allen,A. Ambesi-Impiombato,R. Apweiler,R. N. Aturaliya,T. L. Bailey,M. Bansal,L. Baxter,K. W. Beisel,T. Bersano,H. Bono,A. M. Chalk,K. P. Chiu,V. Choudhary,A. Christoffels,D. R. Clutterbuck,M. L. Crowe,E. Dalla,B. P. Dalrymple,B. de Bono,G. Della Gatta,D. di Bernardo,T. Down,P. Engstrom,M. Fagiolini,G. Faulkner,C. F. Fletcher,T. Fukushima,M. Furuno,S. Futaki,M. Gariboldi,P. Georgii-Hemming,T. R. Gingeras,T. Gojobori,R. E. Green,S. Gustincich,M. Harbers,Y. Hayashi,T. K. Hensch,N. Hirokawa,D. Hill,L. Huminiecki,M. Iacono,K. Ikeo,A. Iwama,T. Ishikawa,M. Jakt,A. Kanapin,M. Katoh,Y. Kawasawa,J. Kelso,H. Kitamura,H. Kitano,G. Kollias,S. P. Krishnan,A. Kruger,S. K. Kummerfeld,I. V. Kurochkin,L. F. Lareau,D. Lazarevic,L. Lipovich,J. Liu,S. Liuni,S. McWilliam,M. Madan Babu,M. Madera,L. Marchionni,H. Matsuda,S. Matsuzawa,H. Miki,F. Mignone,S. Miyake,K. Morris,S. Mottagui-Tabar,N. Mulder,N. Nakano,H. Nakauchi,P. Ng,R. Nilsson,S. Nishiguchi,S. Nishikawa,F. Nori,O. Ohara,Y. Okazaki,V. Orlando,K. C. Pang,W. J. Pavan,G. Pavesi,G. Pesole,N. Petrovsky,S. Piazza,J. Reed,J. F. Reid,B. Z. Ring,M. Ringwald,B. Rost,Y. Ruan,S. L. Salzberg,A. Sandelin,C. Schneider,C. Schonbach,K. Sekiguchi,C. A. Semple,S. Seno,L. Sessa,Y. Sheng,Y. Shibata,H. Shimada,K. Shimada,D. Silva,B. Sinclair,S. Sperling,E. Stupka,K. Sugiura,R.

Sultana,Y. Takenaka,K. Taki,K. Tammoja,S. L. Tan,S. Tang,M. S. Taylor,J. Tegner,S. A. Teichmann,H. R. Ueda,E. van Nimwegen,R. Verardo,C. L. Wei,K. Yagi,H. Yamanishi,E. Zabarovsky,S. Zhu,A. Zimmer,W. Hide,C. Bult,S. M. Grimmond,R. D. Teasdale,E. T. Liu,V. Brusic,J. Quackenbush,C. Wahlestedt,J. S. Mattick,D. A. Hume,C. Kai,D. Sasaki,Y. Tomaru,S. Fukuda,M. Kanamori-Katayama,M. Suzuki,J. Aoki,T. Arakawa,J. Iida,K. Imamura,M. Itoh,T. Kato,H. Kawaji,N. Kawagashira,T. Kawashima,M. Kojima,S. Kondo,H. Konno,K. Nakano,N. Ninomiya,T. Nishio,M. Okada,C. Plessy,K. Shibata,T. Shiraki,S. Suzuki,M. Tagami,K. Waki,A. Watahiki,Y. Okamura-Oho,H. Suzuki,J. Kawai and Y. Hayashizaki. 2005. The transcriptional landscape of the mammalian genome. Science 309, no. 5740: 1559-63.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38, no. 6: 626-35.

Carninci, P., A. Westover, Y. Nishiyama, T. Ohsumi, M. Itoh, S. Nagaoka, N. Sasaki, Y. Okazaki, M. Muramatsu, C. Schneider, and Y. Hayashizaki. 1997. High efficiency selection of full-length cDNA by improved biotinylated cap trapper. DNA Res 4, no. 1: 61-6.

Casares, F. and R. S. Mann. 1998. Control of antennal versus leg development in drosophila. Nature 392, no. 6677: 723-6.

Celniker, S. E., L. A. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, and R. H. Waterston. 2009. Unlocking the secrets of the genome. Nature 459, no. 7249: 927-30.

Celniker, S. E., D. A. Wheeler, B. Kronmiller, J. W. Carlson, A. Halpern, S. Patel, M. Adams, M. Champe, S. P. Dugan, E. Frise, A. Hodgson, R. A. George, R. A. Hoskins, T. Laverty, D. M. Muzny, C. R. Nelson, J. M. Pacleb, S. Park, B. D. Pfeiffer, S. Richards, E. J. Sodergren, R. Svirskas, P. E. Tabor, K. Wan, M. Stapleton, G. G. Sutton, C. Venter, G. Weinstock, S. E. Scherer, E. W. Myers, R. A. Gibbs, and G. M. Rubin. 2002. Finishing a whole-genome shotgun: Release 3 of the drosophila melanogaster euchromatic genome sequence. Genome Biol 3, no. 12: RESEARCH0079.

Che, D., S. Jensen, L. Cai, and J. S. Liu. 2005. BEST: Binding-site estimation suite of tools. Bioinformatics 21, no. 12: 2909-11.

Chen, S. T., H. C. Cheng, D. A. Barbash, and H. P. Yang. 2007. Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in drosophila melanogaster. PLoS Genet 3, no. 7: e107.

Cheng, Y., B. Gvakharia, and P. E. Hardin. 1998. Two alternatively spliced transcripts from the drosophila period gene rescue rhythms having different molecular and behavioral characteristics. Mol Cell Biol 18, no. 11: 6505-14.

Chern, T. M., N. Paul, E. van Nimwegen, and M. Zavolan. 2008. Computational analysis of full-length cdnas reveals frequent coupling between transcriptional and splicing programs. DNA Res 15, no. 2: 63-72.

Chernukhin, I., S. Shamsuddin, S. Y. Kang, R. Bergstrom, Y. W. Kwon, W. Yu, J. Whitehead, R. Mukhopadhyay, F. Docquier, D. Farrar, I. Morrison, M. Vigneron, S. Y. Wu, C. M. Chiang, D. Loukinov, V. Lobanenkov, R. Ohlsson, and E. Klenova. 2007. Ctcf interacts with and recruits the largest subunit of RNA polymerase II to ctcf target sites genome-wide. Mol Cell Biol 27, no. 5: 1631-48.

Chintapalli, V. R., J. Wang, and J. A. Dow. 2007. Using flyatlas to identify better drosophila melanogaster models of human disease. Nat Genet 39, no. 6: 715-20.

Churchill, F. B. 1974. William johannsen and the genotype concept. J Hist Biol 7, no. 1: 5-30.

Clark, A. G.,M. B. Eisen,D. R. Smith,C. M. Bergman,B. Oliver,T. A. Markow,T. C. Kaufman,M. Kellis,W. Gelbart,V. N. Iyer,D. A. Pollard,T. B. Sackton,A. M. Larracuente,N. D. Singh,J. P. Abad,D. N. Abt,B. Adryan,M. Aguade,H. Akashi,W. W. Anderson,C. F. Aquadro,D. H. Ardell,R. Arguello,C. G. Artieri,D. A. Barbash,D. Barker,P. Barsanti,P. Batterham,S. Batzoglou,D. Begun,A. Bhutkar,E. Blanco,S. A. Bosak,R. K. Bradley,A. D. Brand,M. R. Brent,A. N. Brooks,R. H. Brown,R. K. Butlin,C. Caggese,B. R. Calvi,A. Bernardo de Carvalho,A. Caspi,S. Castrezana,S. E. Celniker,J. L. Chang,C. Chapple,S. Chatterji,A. Chinwalla,A. Civetta,S. W. Clifton,J. M. Comeron,J. C. Costello,J. A. Coyne,J. Daub,R. G. David,A. L. Delcher,K. Delehaunty,C. B. Do,H. Ebling,K. Edwards,T. Eickbush,J. D. Evans,A. Filipski,S. Findeiss,E. Freyhult,L. Fulton,R. Fulton,A. C. Garcia,A. Gardiner,D. A. Garfield,B. E. Garvin,G. Gibson,D. Gilbert,S. Gnerre,J. Godfrey,R. Good,V. Gotea,B. Gravely,A. J. Greenberg,S. Griffiths-Jones,S. Gross,R. Guigo,E. A. Gustafson,W. Haerty,M. W. Hahn,D. L. Halligan,A. L. Halpern,G. M. Halter,M. V. Han,A. Heger,L. Hillier,A. S. Hinrichs,I. Holmes,R. A. Hoskins,M. J. Hubisz,D. Hultmark,M. A. Huntley,D. B. Jaffe,S. Jagadeeshan,W. R. Jeck,J. Johnson,C. D. Jones,W. C. Jordan,G. H. Karpen,E. Kataoka,P. D. Keightley,P. Kheradpour,E. F. Kirkness,L. B. Koerich,K. Kristiansen,D. Kudrna,R. J. Kulathinal,S. Kumar,R. Kwok,E. Lander,C. H. Langley,R. Lapoint,B. P. Lazzaro,S. J. Lee,L. Levesque,R. Li,C. F. Lin,M. F. Lin,K. Lindblad-Toh,A. Llopart,M. Long,L. Low,E. Lozovsky,J. Lu,M. Luo,C. A. Machado,W. Makalowski,M. Marzo,M. Matsuda,L. Matzkin,B. McAllister,C. S. McBride,B. McKernan,K. McKernan,M. Mendez-Lago,P. Minx,M. U. Mollenhauer,K. Montooth,S. M. Mount,X. Mu,E. Myers,B. Negre,S. Newfeld,R. Nielsen,M. A. Noor,P. O'Grady,L. Pachter,M. Papaceit,M. J. Parisi,M. Parisi,L. Parts,J. S. Pedersen,G. Pesole,A. M. Phillippy,C. P. Ponting,M. Pop,D. Porcelli,J. R. Powell,S. Prohaska,K. Pruitt,M. Puig,H. Quesneville,K. R. Ram,D. Rand,M. D. Rasmussen,L. K. Reed,R. Reenan,A. Reily,K. A. Remington,T. T. Rieger,M. G. Ritchie,C. Robin,Y. H. Rogers,C. Rohde,J. Rozas,M. J. Rubenfield,A. Ruiz,S. Russo,S. L. Salzberg,A. Sanchez-Gracia,D. J. Saranga,H. Sato,S. W. Schaeffer,M. C. Schatz,T. Schlenke,R. Schwartz,C. Segarra,R. S.

Singh,L. Sirot,M. Sirota,N. B. Sisneros,C. D. Smith,T. F. Smith,J. Spieth,D. E. Stage,A. Stark,W. Stephan,R. L. Strausberg,S. Strempel,D. Sturgill,G. Sutton,G. G. Sutton,W. Tao,S. Teichmann,Y. N. Tobari,Y. Tomimura,J. M. Tsolas,V. L. Valente,E. Venter,J. C. Venter,S. Vicario,F. G. Vieira,A. J. Vilella,A. Villasante,B. Walenz,J. Wang,M. Wasserman,T. Watts,D. Wilson,R. K. Wilson,R. A. Wing,M. F. Wolfner,A. Wong,G. K. Wong,C. I. Wu,G. Wu,D. Yamamoto,H. P. Yang,S. P. Yang,J. A. Yorke,K. Yoshida,E. Zdobnov,P. Zhang,Y. Zhang,A. V. Zimin,J. Baldwin,A. Abdouelleil,J. Abdulkadir,A. Abebe,B. Abera,J. Abreu,S. C. Acer,L. Aftuck,A. Alexander,P. An,E. Anderson,S. Anderson,H. Arachi,M. Azer,P. Bachantsang,A. Barry,T. Bayul,A. Berlin,D. Bessette,T. Bloom,J. Blye,L. Boguslavskiy,C. Bonnet,B. Boukhgalter,I. Bourzgui,A. Brown,P. Cahill,S. Channer,Y. Cheshatsang,L. Chuda,M. Citroen,A. Collymore,P. Cooke,M. Costello,K. D'Aco,R. Daza,G. De Haan,S. DeGray,C. DeMaso,N. Dhargay,K. Dooley,E. Dooley,M. Doricent,P. Dorje,K. Dorjee,A. Dupes,R. Elong,J. Falk,A. Farina,S. Faro,D. Ferguson,S. Fisher,C. D. Foley,A. Franke,D. Friedrich,L. Gadbois,G. Gearin,C. R. Gearin,G. Giannoukos,T. Goode,J. Graham,E. Grandbois,S. Grewal,K. Gyaltsen,N. Hafez,B. Hagos,J. Hall,C. Henson,A. Hollinger,T. Honan,M. D. Huard,L. Hughes,B. Hurhula,M. E. Husby,A. Kamat,B. Kanga,S. Kashin,D. Khazanovich,P. Kisner,K. Lance,M. Lara,W. Lee,N. Lennon,F. Letendre,R. LeVine,A. Lipovsky,X. Liu,J. Liu,S. Liu,T. Lokyitsang,Y. Lokyitsang,R. Lubonja,A. Lui,P. MacDonald,V. Magnisalis,K. Maru,C. Matthews,W. McCusker,S. McDonough,T. Mehta,J. Meldrim,L. Meneus,O. Mihai,A. Mihalev,T. Mihova,R. Mittelman,V. Mlenga,A. Montmayeur,L. Mulrain,A. Navidi,J. Naylor,T. Negash,T. Nguyen,N. Nguyen,R. Nicol,C. Norbu,N. Norbu,N. Novod,B. O'Neill,S. Osman,E. Markiewicz,O. L. Oyono,C. Patti,P. Phunkhang,F. Pierre,M. Priest,S. Raghuraman,F. Rege,R. Reyes,C. Rise,P. Rogov,K. Ross,E. Ryan,S. Settipalli,T. Shea,N. Sherpa,L. Shi,D. Shih,T. Sparrow,J. Spaulding,J. Stalker,N. Stange-Thomann,S. Stavropoulos,C. Stone,C. Strader,S. Tesfaye,T. Thomson,Y. Thoulutsang,D. Thoulutsang,K. Topham,I. Topping,T. Tsamla,H. Vassiliev,A. Vo,T. Wangchuk,T. Wangdi,M. Weiand,J. Wilkinson,A. Wilson,S. Yadav,G. Young,Q. Yu,L. Zembek,D. Zhong,A. Zimmer,Z. Zwirko,P. Alvarez,W. Brockman,J. Butler,C. Chin,M. Grabherr,M. Kleber,E. Mauceli and I. MacCallum. 2007. Evolution of genes and genomes on the drosophila phylogeny. Nature 450, no. 7167: 203-18.

Corbin, V. and T. Maniatis. 1989. The role of specific enhancer-promoter interactions in the drosophila adh promoter switch. Genes Dev 3, no. 12B: 2191-20.

Core, L. J., J. J. Waterfall, and J. T. Lis. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, no. 5909: 1845-8.

Cramer, P., J. F. Caceres, D. Cazalla, S. Kadener, A. F. Muro, F. E. Baralle, and A. R. Kornblihtt. 1999. Coupling of transcription with alternative splicing: RNA pol II promoters modulate sf2/asf and 9g8 effects on an exonic splicing enhancer. Mol Cell 4, no. 2: 251-8.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. Weblogo: A sequence logo generator. Genome Res 14, no. 6: 1188-90.

Davuluri, R. V., Y. Suzuki, S. Sugano, C. Plass, and T. H. Huang. 2008. The functional consequences of alternative promoter use in mammalian genomes. Trends Genet 24, no. 4: 167-77.

De Renzis, S., O. Elemento, S. Tavazoie, and E. F. Wieschaus. 2007. Unmasking activation of the zygotic genome using chromosomal deletions in the drosophila embryo. PLoS Biol 5, no. 5: e117.

Deato, M. D. and R. Tjian. 2007. Switching of the core transcription machinery during myogenesis. Genes Dev 21, no. 17: 2137-49.

Down, T. A., C. M. Bergman, J. Su, and T. J. Hubbard. 2007. Large-scale discovery of promoter motifs in drosophila melanogaster. PLoS Comput Biol 3, no. 1: e7.

Dynlacht, B. D., T. Hoey, and R. Tjian. 1991. Isolation of coactivators associated with the tata-binding protein that mediate transcriptional activation. Cell 66, no. 3: 563-76.

Engstrom, P. G., S. J. Ho Sui, O. Drivenes, T. S. Becker, and B. Lenhard. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res 17, no. 12: 1898-908.

Esteban, J. A., M. Salas, and L. Blanco. 1993. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. J Biol Chem 268, no. 4: 2719-26.

Fejes-Toth, K., V. Sotirova, R. Sachidanandam, G. Assaf, G.J. Hannon, P. Kapranov, S. Foissac, A.T. Willingham, R. Duttagupta, E. Dumais, and T.R. Gingeras. 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short rnas. Nature 457, no. 7232: 1028-32.

FitzGerald, P. C., D. Sturgill, A. Shyakhtenko, B. Oliver, and C. Vinson. 2006. Comparative genomics of drosophila and human core promoters. Genome Biol 7, no. 7: R53.

Frith, M. C., J. Ponjavic, D. Fredman, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and A. Sandelin. 2006. Evolutionary turnover of mammalian transcription start sites. Genome Res 16, no. 6: 713-22.

Frith, M. C., E. Valen, A. Krogh, Y. Hayashizaki, P. Carninci, and A. Sandelin. 2008. A code for transcription initiation in mammalian genomes. Genome Res 18, no. 1: 1-12.

Fu, Y., M. Sinha, C. L. Peterson, and Z. Weng. 2008. The insulator binding protein ctcf positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet 4, no. 7: e1000138.

Fujii, S. and H. Amrein. 2002. Genes expressed in the drosophila head reveal a role for fat cells in sex-specific physiology. Embo J 21, no. 20: 5353-63.

Gazdag, E., A. Santenard, C. Ziegler-Birling, G. Altobelli, O. Poch, L. Tora, and M. E. Torres-Padilla. 2009. Tbp2 is essential for germ cell development by regulating transcription and chromatin condensation in the oocyte. Genes Dev 23, no. 18: 2210-23.

Gelfand, A. E. and P. Vounatsou. 2003. Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics 4, no. 1: 11-25.

Gendra, E., D. F. Colgan, B. Meany, and M. M. Konarska. 2007. A sequence motif in the simian virus 40 (sv40) early core promoter affects alternative splicing of transcribed mRNA. J Biol Chem 282, no. 16: 11648-57.

Gibson, G. and S.V. Muse. 2002. A primer of genome science: Sinauer Associates Inc.

Gorski, S. M., S. Chittaranjan, E. D. Pleasance, J. D. Freeman, C. L. Anderson, R. J. Varhol, S. M. Coughlin, S. D. Zuyderduyn, S. J. Jones, and M. A. Marra. 2003. A SAGE approach to discovery of genes involved in autophagic cell death. Curr Biol 13, no. 4: 358-63.

Green, M. R. 2000. Tbp-associated factors (tafiis): Multiple, selective transcriptional mediators in common complexes. Trends Biochem Sci 25, no. 2: 59-63.

Gross, P. and T. Oelgeschlager. 2006. Core promoter-selective RNA polymerase II transcription. Biochem Soc Symp, no. 73: 225-36.

Gunderson, F. Q. and T. L. Johnson. 2009. Acetylation by the transcriptional coactivator gcn5 plays a novel role in co-transcriptional spliceosome assembly. PLoS Genet 5, no. 10: e1000682.

Hendrix, D. A., J. W. Hong, J. Zeitlinger, D. S. Rokhsar, and M. S. Levine. 2008. Promoter elements associated with RNA pol II stalling in the drosophila embryo. Proc Natl Acad Sci U S A 105, no. 22: 7762-7.

Henikoff, S. and K. Ahmad. 2005. Assembly of variant histones into chromatin. Annu Rev Cell Dev Biol 21: 133-53.

Hernandez, G., P. Vazquez-Pianzola, J. M. Sierra, and R. Rivera-Pomar. 2004. Internal ribosome entry site drives cap-independent translation of reaper and heat shock protein 70 mRNAs in drosophila embryos. RNA 10, no. 11: 1783-97.

Hertz, G. Z. and G. D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, no. 7-8: 563-77.

Hiller, M. A., T. Y. Lin, C. Wood, and M. T. Fuller. 2001. Developmental regulation of transcription by a tissue-specific taf homolog. Genes Dev 15, no. 8: 1021-30.

Hiller, M., X. Chen, M. J. Pringle, M. Suchorolski, Y. Sancak, S. Viswanathan, B. Bolival, T. Y. Lin, S. Marino, and M. T. Fuller. 2004. Testis-specific taf homologs collaborate to control a tissue-specific transcription program. Development 131, no. 21: 5297-308.

Hochheimer, A. and R. Tjian. 2003. Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. Genes Dev 17, no. 11: 1309-20.

Hochheimer, A., S. Zhou, S. Zheng, M. C. Holmes, and R. Tjian. 2002. Trf2 associates with dref and directs promoter-selective gene expression in drosophila. Nature 420, no. 6914: 439-45.

Holmes, M. C. and R. Tjian. 2000. Promoter-selective properties of the tbp-related factor trf1. Science 288, no. 5467: 867-70.

Hosono, S., A. F. Faruqi, F. B. Dean, Y. Du, Z. Sun, X. Wu, J. Du, S. F. Kingsmore, M. Egholm, and R. S. Lasken. 2003. Unbiased whole-genome amplification directly from clinical samples. Genome Res 13, no. 5: 954-64.

Hsu, J. Y., T. Juven-Gershon, M. T. Marr, 2nd, K. J. Wright, R. Tjian, and J. T. Kadonaga. 2008. Tbp, mot1, and nc2 establish a regulatory circuit that controls dpe-dependent versus tata-dependent transcription. Genes Dev 22, no. 17: 2353-8.

Huang, W., J. R. Nevins, and U. Ohler. 2007. Phylogenetic simulation of promoter evolution: Estimation and modeling of binding site turnover events and assessment of their impact on alignment tools. Genome Biol 8, no. 10: R225.

Hughes, J. D., P. W. Estep, S. Tavazoie, and G. M. Church. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. J Mol Biol 296, no. 5: 1205-14.

Isogai, Y., S. Keles, M. Prestel, A. Hochheimer, and R. Tjian. 2007. Transcription of histone gene cluster by differential core-promoter factors. Genes Dev 21, no. 22: 2936-49.

Jenuwein, T. and C. D. Allis. 2001. Translating the histone code. Science 293, no. 5532: 1074-80.

Jin, C., C. Zang, G. Wei, K. Cui, W. Peng, K. Zhao, and G. Felsenfeld. 2009. H3.3/h2a.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. Nat Genet 41, no. 8: 941-5.

Juven-Gershon, T., S. Cheng, and J. T. Kadonaga. 2006. Rational design of a super core promoter that enhances gene expression. Nat Methods 3, no. 11: 917-22.

Juven-Gershon, T., J. Y. Hsu, J. W. Theisen, and J. T. Kadonaga. 2008. The RNA polymerase II core promoter - the gateway to transcription. Curr Opin Cell Biol 20, no. 3: 253-9.

Juven-Gershon, T. and J. T. Kadonaga. 2009. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev Biol.

Kampa, D., J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana, and T. R. Gingeras. 2004. Novel rnas

identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14, no. 3: 331-42.

Kapranov, P., J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermuller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammana, and T. R. Gingeras. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316, no. 5830: 1484-8.

Katsani, K. R., M. A. Hajibagheri, and C. P. Verrijzer. 1999. Co-operative DNA binding by gaga transcription factor requires the conserved btb/poz domain and reorganizes promoter topology. Embo J 18, no. 3: 698-708.

Kawaji, H., M. C. Frith, S. Katayama, A. Sandelin, C. Kai, J. Kawai, P. Carninci, and Y. Hayashizaki. 2006. Dynamic usage of transcription start sites within core promoters. Genome Biol 7, no. 12: R118.

Kawaji, H., J. Severin, M. Lizio, A. Waterhouse, S. Katayama, K. M. Irvine, D. A. Hume, A. R. Forrest, H. Suzuki, P. Carninci, Y. Hayashizaki, and C. O. Daub. 2009. The fantom web resource: From mammalian transcriptional landscape to its dynamic regulation. Genome Biol 10, no. 4: R40.

Kimura, K., A. Wakamatsu, Y. Suzuki, T. Ota, T. Nishikawa, R. Yamashita, J. Yamamoto, M. Sekine, K. Tsuritani, H. Wakaguri, S. Ishii, T. Sugiyama, K. Saito, Y. Isono, R. Irie, N. Kushida, T. Yoneyama, R. Otsuka, K. Kanda, T. Yokoi, H. Kondo, M. Wagatsuma, K. Murakawa, S. Ishida, T. Ishibashi, A. Takahashi-Fujii, T. Tanase, K. Nagai, H. Kikuchi, K. Nakai, T. Isogai, and S. Sugano. 2006. Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. Genome Res 16, no. 1: 55-65.

Klipper-Aurbach, Y., M. Wasserman, N. Braunspiegel-Weintrob, D. Borstein, S. Peleg, S. Assa, M. Karp, Y. Benjamini, Y. Hochberg, and Z. Laron. 1995. Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses 45, no. 5: 486-90.

Kodzius, R., M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, Y. Hayashizaki, and P. Carninci. 2006. Cage: Cap analysis of gene expression. Nat Methods 3, no. 3: 211-22.

Koonin, E. V., Y. I. Wolf, and G. P. Karev. 2002. The structure of the protein universe and genome evolution. Nature 420, no. 6912: 218-23.

Krivan, W. and W. W. Wasserman. 2001. A predictive model for regulatory sequences directing liver-specific transcription. Genome Res 11, no. 9: 1559-66.

Kulp, D., D. Haussler, M. G. Reese, and F. H. Eeckman. 1996. A generalized hidden markov model for the recognition of human genes in DNA. Proc Int Conf Intell Syst Mol Biol 4: 134-42.

Kutach, A. K. and J. T. Kadonaga. 2000. The downstream promoter element dpe appears to be as widely used as the tata box in drosophila core promoters. Mol Cell Biol 20, no. 13: 4754-64.

Kwon, S. Y., H. Xiao, B. P. Glover, R. Tjian, C. Wu, and P. Badenhorst. 2008. The nucleosome remodeling factor (nurf) regulates genes involved in drosophila innate immunity. Dev Biol 316, no. 2: 538-47.

Kwon, S. Y., H. Xiao, C. Wu, and P. Badenhorst. 2009. Alternative splicing of nurf301 generates distinct nurf chromatin remodeling complexes with altered modified histone binding specificities. PLoS Genet 5, no. 7: e1000574.

Lander, E. S.,L. M. Linton,B. Birren,C. Nusbaum,M. C. Zody,J. Baldwin,K. Devon,K. Dewar,M. Doyle,W. FitzHugh,R. Funke,D. Gage,K. Harris,A. Heaford,J. Howland,L. Kann,J. Lehoczky,R. LeVine,P. McEwan,K. McKernan,J. Meldrim,J. P. Mesirov,C. Miranda,W. Morris,J. Naylor,C. Raymond,M. Rosetti,R. Santos,A. Sheridan,C. Sougnez,N. Stange-Thomann,N. Stojanovic,A. Subramanian,D. Wyman,J. Rogers,J. Sulston,R. Ainscough,S. Beck,D. Bentley,J. Burton,C. Clee,N. Carter,A. Coulson,R. Deadman,P. Deloukas,A. Dunham,I. Dunham,R. Durbin,L. French,D. Grafham,S. Gregory,T. Hubbard,S. Humphray,A. Hunt,M. Jones,C. Lloyd,A. McMurray,L. Matthews,S. Mercer,S. Milne,J. C. Mullikin,A. Mungall,R. Plumb,M. Ross,R. Shownkeen,S. Sims,R. H. Waterston,R. K. Wilson,L. W. Hillier,J. D. McPherson,M. A. Marra,E. R. Mardis,L. A. Fulton,A. T. Chinwalla,K. H. Pepin,W. R. Gish,S. L. Chissoe,M. C. Wendl,K. D. Delehaunty,T. L. Miner,A. Delehaunty,J. B. Kramer,L. L. Cook,R. S. Fulton,D. L. Johnson,P. J. Minx,S. W. Clifton,T. Hawkins,E. Branscomb,P. Predki,P. Richardson,S. Wenning,T. Slezak,N. Doggett,J. F. Cheng,A. Olsen,S. Lucas,C. Elkin,E. Uberbacher,M. Frazier,R. A. Gibbs,D. M. Muzny,S. E. Scherer,J. B. Bouck,E. J. Sodergren,K. C. Worley,C. M. Rives,J. H. Gorrell,M. L. Metzker,S. L. Naylor,R. S. Kucherlapati,D. L. Nelson,G. M. Weinstock,Y. Sakaki,A. Fujiyama,M. Hattori,T. Yada,A. Toyoda,T. Itoh,C. Kawagoe,H. Watanabe,Y. Totoki,T. Taylor,J. Weissenbach,R. Heilig,W. Saurin,F. Artiguenave,P. Brottier,T. Bruls,E. Pelletier,C. Robert,P. Wincker,D. R. Smith,L. Doucette-Stamm,M. Rubenfield,K. Weinstock,H. M. Lee,J. Dubois,A. Rosenthal,M. Platzer,G. Nyakatura,S. Taudien,A. Rump,H. Yang,J. Yu,J. Wang,G. Huang,J. Gu,L. Hood,L. Rowen,A. Madan,S. Qin,R. W. Davis,N. A. Federspiel,A. P. Abola,M. J. Proctor,R. M. Myers,J. Schmutz,M. Dickson,J. Grimwood,D. R. Cox,M. V. Olson,R. Kaul,C. Raymond,N. Shimizu,K. Kawasaki,S. Minoshima,G. A. Evans,M. Athanasiou,R. Schultz,B. A. Roe,F. Chen,H. Pan,J. Ramser,H. Lehrach,R. Reinhardt,W. R. McCombie,M. de la Bastide,N. Dedhia,H. Blocker,K. Hornischer,G. Nordsiek,R. Agarwala,L. Aravind,J. A. Bailey,A. Bateman,S. Batzoglou,E. Birney,P. Bork,D. G. Brown,C. B. Burge,L. Cerutti,H. C. Chen,D. Church,M. Clamp,R. R. Copley,T. Doerks,S. R. Eddy,E. E. Eichler,T. S. Furey,J. Galagan,J. G. Gilbert,C. Harmon,Y. Hayashizaki,D. Haussler,H. Hermjakob,K. Hokamp,W. Jang,L. S. Johnson,T. A. Jones,S. Kasif,A. Kaspryzk,S. Kennedy,W. J. Kent,P. Kitts,E. V. Koonin,I. Korf,D. Kulp,D. Lancet,T. M. Lowe,A. McLysaght,T.

Mikkelsen,J. V. Moran,N. Mulder,V. J. Pollara,C. P. Ponting,G. Schuler,J. Schultz,G. Slater,A. F. Smit,E. Stupka,J. Szustakowski,D. Thierry-Mieg,J. Thierry-Mieg,L. Wagner,J. Wallis,R. Wheeler,A. Williams,Y. I. Wolf,K. H. Wolfe,S. P. Yang,R. F. Yeh,F. Collins,M. S. Guyer,J. Peterson,A. Felsenfeld,K. A. Wetterstrand,A. Patrinos,M. J. Morgan,P. de Jong,J. J. Catanese,K. Osoegawa,H. Shizuya,S. Choi and Y. J. Chen. 2001. Initial sequencing and analysis of the human genome. Nature 409, no. 6822: 860-921.

Landry, J. R., D. L. Mager, and B. T. Wilhelm. 2003. Complex controls: The role of alternative promoters in mammalian genomes. Trends Genet 19, no. 11: 640-8.

Latchman, D. 2005. Gene regulation: Taylor and Francis Group.

Lee, A. M. and C. T. Wu. 2006. Enhancer-promoter communication at the yellow gene of drosophila melanogaster: Diverse promoters participate in and regulate trans interactions. Genetics 174, no. 4: 1867-80.

Lee, C., X. Li, A. Hechmer, M. Eisen, M. D. Biggin, B. J. Venters, C. Jiang, J. Li, B. F. Pugh, and D. S. Gilmour. 2008. Nelf and gaga factor are linked to promoter-proximal pausing at many genes in drosophila. Mol Cell Biol 28, no. 10: 3290-300.

Lee, D. H., N. Gershenzon, M. Gupta, I. P. Ioshikhes, D. Reinberg, and B. A. Lewis. 2005. Functional characterization of core promoter elements: The downstream core element is recognized by taf1. Mol Cell Biol 25, no. 21: 9674-86.

Lee, W., D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow. 2007. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 39, no. 10: 1235-44.

Liang, H., Y. S. Lin, and W. H. Li. 2008. Fast evolution of core promoters in primate genomes. Mol Biol Evol 25, no. 6: 1239-44.

Lifton, R. P., M. L. Goldberg, R. W. Karp, and D. S. Hogness. 1978. The organization of the histone genes in drosophila melanogaster: Functional and evolutionary implications. Cold Spring Harb Symp Quant Biol 42 Pt 2: 1047-51.

Lis, J. 1998. Promoter-associated pausing in promoter architecture and postinitiation transcriptional regulation. Cold Spring Harb Symp Quant Biol 63: 347-56.

Littlefield, O., Y. Korkhin, and P. B. Sigler. 1999. The structural basis for the oriented assembly of a tbp/tfb/promoter complex. Proc Natl Acad Sci U S A 96, no. 24: 13668-73.

Liu, H., J. K. Jang, J. Graham, K. Nycz, and K. S. McKim. 2000. Two genes required for meiotic recombination in drosophila are expressed from a dicistronic message. Genetics 154, no. 4: 1735-46.

Lo, K. and S. T. Smale. 1996. Generality of a functional initiator consensus sequence. Gene 182, no. 1-2: 13-22.

Loseva, O. and Y. Engstrom. 2004. Analysis of signal-dependent changes in the proteome of drosophila blood cells during an immune response. Mol Cell Proteomics 3, no. 8: 796-808.

Ma, C., M. Lyons-Weiler, W. Liang, W. LaFramboise, J. R. Gilbertson, M. J. Becich, and F. A. Monzon. 2006. In vitro transcription amplification and labeling methods contribute to the variability of gene expression profiling with DNA microarrays. J Mol Diagn 8, no. 2: 183-92.

MacIsaac, K. D. and E. Fraenkel. 2006. Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol 2, no. 4: e36.

Magyar, A., E. Bakos, and A. Varadi. 1995. Structure and tissue-specific expression of the drosophila melanogaster organellar-type ca(2+)-atpase gene. Biochem J 310 (Pt 3): 757-63.

Mahmoudi, T., K. R. Katsani, and C. P. Verrijzer. 2002. Gaga can mediate enhancer function in trans by linking two separate DNA molecules. Embo J 21, no. 7: 1775-81.

Mahony, S. and P. V. Benos. 2007. Stamp: A web tool for exploring DNA-binding motif similarities. Nucleic Acids Res 35, no. Web Server issue: W253-8.

Majoros, W.H. 2007. Methods for computational gene prediction: Cambridge University Press.

Manak, J. R., S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, A. Piccolboni, and T. R. Gingeras. 2006. Biological function of unannotated transcription during the early development of drosophila melanogaster. Nat Genet 38, no. 10: 1151-8.

Margolis, J. S., M. L. Borowsky, E. Steingrimsson, C. W. Shim, J. A. Lengyel, and J. W. Posakony. 1995. Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element. Development 121, no. 9: 3067-77.

Margolis, J. S., M. Borowsky, C. W. Shim, and J. W. Posakony. 1994. A small region surrounding the distal promoter of the hunchback gene directs maternal expression. Dev Biol 163, no. 2: 381-8.

Martens, J. A., L. Laprade, and F. Winston. 2004. Intergenic transcription is required to repress the saccharomyces cerevisiae ser3 gene. Nature 429, no. 6991: 571-4.

Marx, J. 2000. Cell biology. New clues to how genes are controlled. Science 290, no. 5494: 1066-7.

Mavrich, T. N., C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, I. Albert, and B. F. Pugh. 2008. Nucleosome organization in the drosophila genome. Nature 453, no. 7193: 358-62.

Megraw, M., F. Pereira, S. T. Jensen, U. Ohler, and A. G. Hatzigeorgiou. 2009. A transcription factor affinity-based code for mammalian transcription initiation. Genome Res 19, no. 4: 644-56.

Metcalf, C. E. and D. A. Wassarman. 2006. DNA binding properties of taf1 isoforms with two at-hooks. J Biol Chem 281, no. 40: 30015-23.

_____. 2007. Nucleolar colocalization of taf1 and testis-specific tafs during drosophila spermatogenesis. Dev Dyn 236, no. 10: 2836-43.

Min, K. T. and S. Benzer. 1999. Preventing neurodegeneration in the drosophila mutant bubblegum. Science 284, no. 5422: 1985-8.

Misra, S., M. A. Crosby, C. J. Mungall, B. B. Matthews, K. S. Campbell, P. Hradecky, Y. Huang, J. S. Kaminker, G. H. Millburn, S. E. Prochnik, C. D. Smith, J. L. Tupy, E. J. Whitfied, L. Bayraktaroglu, B. P. Berman, B. R. Bettencourt, S. E. Celniker, A. D. de Grey, R. A. Drysdale, N. L. Harris, J. Richter, S. Russo, A. J. Schroeder, S. Q. Shu, M. Stapleton, C. Yamada, M. Ashburner, W. M. Gelbart, G. M. Rubin, and S. E. Lewis. 2002. Annotation of the drosophila melanogaster euchromatic genome: A systematic review. Genome Biol 3, no. 12: RESEARCH0083.

Mito, Y., J. G. Henikoff, and S. Henikoff. 2005. Genome-scale profiling of histone h3.3 replacement patterns. Nat Genet 37, no. 10: 1090-7.

_____. 2007. Histone replacement marks the boundaries of cis-regulatory domains. Science 315, no. 5817: 1408-11.

Mizuguchi, G., A. Vassilev, T. Tsukiyama, Y. Nakatani, and C. Wu. 2001. Atp-dependent nucleosome remodeling and histone hyperacetylation synergistically facilitate transcription of chromatin. J Biol Chem 276, no. 18: 14773-83.

modENCODE. 2010. Accessed. Available from http://www.modencode.org.

Morozova, T. V., R. R. Anholt, and T. F. Mackay. 2006. Transcriptional response to alcohol exposure in drosophila melanogaster. Genome Biol 7, no. 10: R95.

Moses, A. M., D. A. Pollard, D. A. Nix, V. N. Iyer, X. Y. Li, M. D. Biggin, and M. B. Eisen. 2006. Large-scale turnover of functional transcription factor binding sites in drosophila. PLoS Comput Biol 2, no. 10: e130.

Mottus, R. C., I. P. Whitehead, M. O'Grady, R. E. Sobel, R. H. Burr, G. B. Spiegelman, and T. A. Grigliatti. 1997. Unique gene organization: Alternative splicing in drosophila produces two structurally unrelated proteins. Gene 198, no. 1-2: 229-36.

Muse, G. W., D. A. Gilchrist, S. Nechaev, R. Shah, J. S. Parker, S. F. Grissom, J. Zeitlinger, and K. Adelman. 2007. RNA polymerase is poised for activation across the genome. Nat Genet 39, no. 12: 1507-11.

Nata, K., T. Sugimoto, A. Tohgo, T. Takamura, N. Noguchi, A. Matsuoka, T. Numakunai, K. Shikama, H. Yonekura, S. Takasawa, and et al. 1995. The structure of the aplysia kurodai gene encoding adp-ribosyl cyclase, a second-messenger enzyme. Gene 158, no. 2: 213-8.

National-Diagnostics. 2010. Primer extension. Accessed. Available from http://nationaldiagnostics.com/article_info.php/articles_id/29.

_____. 2010. Ribonuclease protection. Accessed. Available from http://nationaldiagnostics.com/article_info.php/articles_id/28.

_____. 2010. S1 mapping. Accessed. Available from http://nationaldiagnostics.com/article_info.php/articles_id/27.

Nechaev, S., D. C. Fargo, G. dos Santos, L. Liu, Y. Gao, and K. Adelman. 2009. Global analysis of short rnas reveals widespread promoter-proximal stalling and arrest of pol II in drosophila. Science 327, no. 5963: 335-8.

Negre, N., C. D. Brown, P. K. Shah, P. Kheradpour, C. A. Morrison, J. G. Henikoff, X. Feng, K. Ahmad, S. Russell, R. A. White, L. Stein, S. Henikoff, M. Kellis, and K. P. White. 2010. A comprehensive map of insulator elements for the drosophila genome. PLoS Genet 6, no. 1: e1000814.

Nelson, C. E., B. M. Hersh, and S. B. Carroll. 2004. The regulatory content of intergenic DNA shapes genome architecture. Genome Biol 5, no. 4: R25.

Ni, T., D.L. Corcoran, E.A. Rach, S. Song, E.P. Spana, Y. Gao, U. Ohler, and J. Zhu. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. In review.

Nobel lectures physiology or medicine 1922-1941. 1965. Amsterdam: Elsevier Publishing Company.

Ohler, U. 2006. Identification of core promoter modules in drosophila and their application in accurate transcription start site prediction. Nucleic Acids Res 34, no. 20: 5943-50.

Ohler, U. and M. Frith. 2005. Models for complex eukaryotic regulatory DNA sequences. In Information processing and living systems: Imperial College Press.

Ohler, U., G. C. Liao, H. Niemann, and G. M. Rubin. 2002. Computational analysis of core promoters in the drosophila genome. Genome Biol 3, no. 12: RESEARCH0087.

Ohler, U., H. Niemann, Gc Liao, and G. M. Rubin. 2001. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics 17 Suppl 1: S199-206.

Ohler, U. and D. A. Wassarman. 2010. Promoting developmental transcription. Development 137, no. 1: 15-26.

Ohtsuki, S. and M. Levine. 1998. Gaga mediates the enhancer blocking activity of the eve promoter in the drosophila embryo. Genes Dev 12, no. 21: 3325-30.

Orphanides, G. and D. Reinberg. 2002. A unified theory of gene expression. Cell 108, no. 4: 439-51.

Parisi, M., R. Nuttall, P. Edwards, J. Minor, D. Naiman, J. Lu, M. Doctolero, M. Vainer, C. Chan, J. Malley, S. Eastman, and B. Oliver. 2004. A survey of ovary-, testis-, and soma-biased gene expression in drosophila melanogaster adults. Genome Biol 5, no. 6: R40.

Pavlidis, P., T. S. Furey, M. Liberto, D. Haussler, and W. N. Grundy. 2001. Promoter region-based classification of genes. Pac Symp Biocomput: 151-63.

Picard, R.R. and R.D. Cook. 1984. Cross-validation of regression models. In 79:575-83: American Statistical Association.

Pierstorff, N., C. M. Bergman, and T. Wiehe. 2006. Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. Bioinformatics 22, no. 23: 2858-64.

Pollard, D. A., A. M. Moses, V. N. Iyer, and M. B. Eisen. 2006. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. BMC Bioinformatics 7: 376.

Ponjavic, J., B. Lenhard, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and A. Sandelin. 2006. Transcriptional and structural impact of tata-initiation site spacing in mammalian core promoters. Genome Biol 7, no. 8: R78.

Purnell, B. A., P. A. Emanuel, and D. S. Gilmour. 1994. Tfiid sequence recognition of the initiator and sequences farther downstream in drosophila class II genes. Genes Dev 8, no. 7: 830-42.

Pyhtila, B., T. Zheng, P. J. Lager, J. D. Keene, M. C. Reedy, and C. V. Nicchitta. 2008. Signal sequence- and translation-independent mRNA localization to the endoplasmic reticulum. Rna 14, no. 3: 445-53.

Rabenstein, M. D., S. Zhou, J. T. Lis, and R. Tjian. 1999. Tata box-binding protein (tbp)-related factor 2 (trf2), a third member of the tbp family. Proc Natl Acad Sci U S A 96, no. 9: 4791-6.

Rach, E. . 2004. Power law distributions of gene family sizes, Cornell University.

Rach, E. A., H. Y. Yuan, W. H. Majoros, P. Tomancak, and U. Ohler. 2009. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the drosophila genome. Genome Biol 10, no. 7: R73.

Raisner, R. M., P. D. Hartley, M. D. Meneghini, M. Z. Bao, C. L. Liu, S. L. Schreiber, O. J. Rando, and H. D. Madhani. 2005. Histone variant h2a.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell 123, no. 2: 233-48.

Reichert, V. L., H. Le Hir, M. S. Jurica, and M. J. Moore. 2002. 5' exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. Genes Dev 16, no. 21: 2778-91.

Remenyi, A., H. R. Scholer, and M. Wilmanns. 2004. Combinatorial control of gene expression. Nat Struct Mol Biol 11, no. 9: 812-5.

RIKEN, Omics Resource Development Unit. 2010. Cap Analysis Gene Expression technology. Accessed. Available from http://www.riken.go.jp/engn/r-world/research/lab/osc/omires/result.html.

Royce, T. E., J. S. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman, M. Snyder, and M. Gerstein. 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends Genet 21, no. 8: 466-75.

Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White. 1998. Microbial gene identification using interpolated markov models. Nucleic Acids Res 26, no. 2: 544-8.

Sandelin, A., P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, and D. A. Hume. 2007. Mammalian RNA polymerase II core promoters: Insights from genome-wide studies. Nat Rev Genet 8, no. 6: 424-36.

Sang, J.H. 2001. Drosophila melanogaster: The fruit fly. In Encyclopedia of Genetics, ed. E.C.R. Reeve:157. Chicago, IL: Fitzroy Dearborn Publishers.

Sawado, T., F. Hirose, Y. Takahashi, T. Sasaki, T. Shinomiya, K. Sakaguchi, A. Matsukage, and M. Yamaguchi. 1998. The DNA replication-related element (dre)/dre-binding factor system is a transcriptional regulator of the drosophila e2f gene. J Biol Chem 273, no. 40: 26042-51.

Schafer, B. 2010. STATCON (Statistische Consulting and Software): Equbits foresight accuracy. Accessed. Available from http://www.statcon.de/statconshop/product_info.htm?products_id=453.

Schedl, P. and J. R. Broach. 2003. Making good neighbors: The right fence for the right job. Nat Struct Biol 10, no. 4: 241-3.

Schier, A. F. 2007. The maternal-zygotic transition: Death and birth of rnas. Science 316, no. 5823: 406-7.

Schmid, C. D., R. Perier, V. Praz, and P. Bucher. 2006. Epd in its twentieth year: Towards complete promoter coverage of selected model organisms. Nucleic Acids Res 34, no. Database issue: D82-5.

Schmid, C. D., V. Praz, M. Delorenzi, R. Perier, and P. Bucher. 2004. The eukaryotic promoter database epd: The impact of in silico primer extension. Nucleic Acids Res 32, no. Database issue: D82-5.

Schmid, C. D., T. Sengstag, P. Bucher, and M. Delorenzi. 2007. Madap, a flexible clustering tool for the interpretation of one-dimensional genome annotation data. Nucleic Acids Res 35, no. Web Server issue: W201-5.

Schneider, T. D. and R. M. Stephens. 1990. Sequence logos: A new way to display consensus sequences. Nucleic Acids Res 18, no. 20: 6097-100.

Schones, D. E., K. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell 132, no. 5: 887-98.

Schug, J., W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert, Jr. 2005. Promoter features related to tissue specificity as measured by shannon entropy. Genome Biol 6, no. 4: R33.

Schwanbeck, R., H. Xiao, and C. Wu. 2004. Spatial contacts and nucleosome step movements induced by the nurf chromatin remodeling complex. J Biol Chem 279, no. 38: 39933-41.

Seila, A. C., J. M. Calabrese, S. S. Levine, G. W. Yeo, P. B. Rahl, R. A. Flynn, R. A. Young, and P. A. Sharp. 2008. Divergent transcription from active promoters. Science 322, no. 5909: 1849-51.

Sharan, R. and E. W. Myers. 2005. A motif-based framework for recognizing sequence families. Bioinformatics 21 Suppl 1: i387-93.

Shibuya, T., T. O. Tange, N. Sonenberg, and M. J. Moore. 2004. Eif4aiii binds spliced mRNA in the exon junction complex and is essential for nonsense-mediated decay. Nat Struct Mol Biol 11, no. 4: 346-51.

Shiraki, T., S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100, no. 26: 15776-81.

Siddharthan, R. 2008. Phylogibbs-mp: Module prediction and discriminative motif-finding by gibbs sampling. PLoS Comput Biol 4, no. 8: e1000156.

Siddharthan, R., E. D. Siggia, and E. van Nimwegen. 2005. Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. PLoS Comput Biol 1, no. 7: e67.

Smale, S. T. and J. T. Kadonaga. 2003. The RNA polymerase II core promoter. Annu Rev Biochem 72: 449-79.

Small, S., R. Kraut, T. Hoey, R. Warrior, and M. Levine. 1991. Transcriptional regulation of a pair-rule stripe in drosophila. Genes Dev 5, no. 5: 827-39.

Smith, M. K. and B. T. Wakimoto. 2007. Complex regulation and multiple developmental functions of misfire, the drosophila melanogaster ferlin gene. BMC Dev Biol 7: 21.

Smith, S. T., P. Wickramasinghe, A. Olson, D. Loukinov, L. Lin, J. Deng, Y. Xiong, J. Rux, R. Sachidanandam, H. Sun, V. Lobanenkov, and J. Zhou. 2009. Genome wide chip-chip analyses reveal important roles for ctcf in drosophila genome organization. Dev Biol 328, no. 2: 518-28.

Song, C.H. and M. Wyse. 2004. Painless gene expression profiling: SAGE (Serial Analysis of Gene Expression). The Science Creative Quarterly, no. 3.

Sonnenburg, S., A. Zien, and G. Ratsch. 2006. Arts: Accurate recognition of transcription starts in human. Bioinformatics 22, no. 14: e472-80.

Spellman, P. T. and G. M. Rubin. 2002. Evidence for large domains of similarly expressed genes in the drosophila genome. J Biol 1, no. 1: 5.

Stapleton, M., G. Liao, P. Brokstein, L. Hong, P. Carninci, T. Shiraki, Y. Hayashizaki, M. Champe, J. Pacleb, K. Wan, C. Yu, J. Carlson, R. George, S. Celniker, and G. M. Rubin. 2002. The drosophila gene collection: Identification of putative full-length cdnas for 70% of d. Melanogaster genes. Genome Res 12, no. 8: 1294-300.

Stark, A., M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S. W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, and M. Kellis. 2007. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. Nature 450, no. 7167: 219-32.

Strachan, T. and A.P. Read. 1999. RACE-PCR facilitates the isolation of 5' and 3' end sequences from cDNA. New York and London: Garland Science.

Suzuki, H.,A. R. Forrest,E. van Nimwegen,C. O. Daub,P. J. Balwierz,K. M. Irvine,T. Lassmann,T. Ravasi,Y. Hasegawa,M. J. de Hoon,S. Katayama,K. Schroder,P. Carninci,Y. Tomaru,M. Kanamori-Katayama,A. Kubosaki,A. Akalin,Y. Ando,E. Arner,M. Asada,H. Asahara,T. Bailey,V. B. Bajic,D. Bauer,A. G. Beckhouse,N. Bertin,J. Bjorkegren,F. Brombacher,E. Bulger,A. M. Chalk,J. Chiba,N. Cloonan,A. Dawe,J. Dostie,P. G. Engstrom,M. Essack,G. J. Faulkner,J. L. Fink,D. Fredman,K. Fujimori,M. Furuno,T. Gojobori,J. Gough,S. M. Grimmond,M. Gustafsson,M. Hashimoto,T. Hashimoto,M. Hatakeyama,S. Heinzel,W. Hide,O. Hofmann,M. Hornquist,L. Huminiecki,K. Ikeo,N. Imamoto,S. Inoue,Y. Inoue,R. Ishihara,T. Iwayanagi,A. Jacobsen,M. Kaur,H. Kawaji,M. C. Kerr,R. Kimura,S. Kimura,Y. Kimura,H. Kitano,H.

Koga,T. Kojima,S. Kondo,T. Konno,A. Krogh,A. Kruger,A. Kumar,B. Lenhard,A. Lennartsson,M. Lindow,M. Lizio,C. Macpherson,N. Maeda,C. A. Maher,M. Maqungo,J. Mar,N. A. Matigian,H. Matsuda,J. S. Mattick,S. Meier,S. Miyamoto,E. Miyamoto-Sato,K. Nakabayashi,Y. Nakachi,M. Nakano,S. Nygaard,T. Okayama,Y. Okazaki,H. Okuda-Yabukami,V. Orlando,J. Otomo,M. Pachkov,N. Petrovsky,C. Plessy,J. Quackenbush,A. Radovanovic,M. Rehli,R. Saito,A. Sandelin,S. Schmeier,C. Schonbach,A. S. Schwartz,C. A. Semple,M. Sera,J. Severin,K. Shirahige,C. Simons,G. St Laurent,M. Suzuki,T. Suzuki,M. J. Sweet,R. J. Taft,S. Takeda,Y. Takenaka,K. Tan,M. S. Taylor,R. D. Teasdale,J. Tegner,S. Teichmann,E. Valen,C. Wahlestedt,K. Waki,A. Waterhouse,C. A. Wells,O. Winther,L. Wu,K. Yamaguchi,H. Yanagawa,J. Yasuda,M. Zavolan,D. A. Hume,T. Arakawa,S. Fukuda,K. Imamura,C. Kai,A. Kaiho,T. Kawashima,C. Kawazu,Y. Kitazume,M. Kojima,H. Miura,K. Murakami,M. Murata,N. Ninomiya,H. Nishiyori,S. Noma,C. Ogawa,T. Sano,C. Simon,M. Tagami,Y. Takahashi,J. Kawai and Y. Hayashizaki. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet 41, no. 5: 553-62.

Suzuki, Y. and S. Sugano. 2003. Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. Methods Mol Biol 221: 73-91.

Suzuki, Y., R. Yamashita, K. Nakai, and S. Sugano. 2002. Dbtss: Database of human transcriptional start sites and full-length cdnas. Nucleic Acids Res 30, no. 1: 328-31.

The development of drosophila melanogaster. 1993. Edited by M Bate and A.M. Arias: Cold Spring Harbor Laboratory Press.

The yeast genome directory. 1997. Nature 387, no. 6632 Suppl: 5.

Tirosh, I. and N. Barkai. 2008. Two strategies for gene regulation by promoter nucleosomes. Genome Res 18, no. 7: 1084-91.

Tirosh, I., N. Barkai, and K. J. Verstrepen. 2009. Promoter architecture and the evolvability of gene expression. J Biol 8, no. 11: 95.

Tolstorukov, M. Y., P. V. Kharchenko, J. A. Goldman, R. E. Kingston, and P. J. Park. 2009. Comparative analysis of h2a.Z nucleosome organization in the human and yeast genomes. Genome Res 19, no. 6: 967-77.

Tomancak, P., A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. 2002. Systematic determination of patterns of gene expression during drosophila embryogenesis. Genome Biol 3, no. 12: RESEARCH0088.

Tomancak, P., B. P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. E. Celniker, and G. M. Rubin. 2007. Global analysis of patterns of gene expression during drosophila embryogenesis. Genome Biol 8, no. 7: R145.

Trifonov, E. N. and J. L. Sussman. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci U S A 77, no. 7: 3816-20.

Tsuchihara, K., Y. Suzuki, H. Wakaguri, T. Irie, K. Tanimoto, S. Hashimoto, K. Matsushima, J. Mizushima-Sugano, R. Yamashita, K. Nakai, D. Bentley, H. Esumi, and S. Sugano. 2009. Massive transcriptional start site analysis of human genes in hypoxia cells. Nucleic Acids Res 37, no. 7: 2249-63.

Tsukiyama, T., P. B. Becker, and C. Wu. 1994. Atp-dependent nucleosome disruption at a heat-shock promoter mediated by binding of gaga transcription factor. Nature 367, no. 6463: 525-32.

Tsuritani, K., T. Irie, R. Yamashita, Y. Sakakibara, H. Wakaguri, A. Kanai, J. Mizushima-Sugano, S. Sugano, K. Nakai, and Y. Suzuki. 2007. Distinct class of putative "Non-conserved" Promoters in humans: Comparative studies of alternative promoters of human and mouse genes. Genome Res 17, no. 7: 1005-14.

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Seal, and H. Zhang. 2009. Flybase: Enhancing drosophila gene ontology annotations. Nucleic Acids Res 37, no. Database issue: D555-9.

Tweedie, S., H. H. Ng, A. L. Barlow, B. M. Turner, B. Hendrich, and A. Bird. 1999. Vestiges of a DNA methylation system in drosophila melanogaster? Nat Genet 23, no. 4: 389-90.

Unicellular organisms: Genomes. 2010. Accessed. Available from http://universe-review.ca/F11-monocell.htm.

Valen, E., G. Pascarella, A. Chalk, N. Maeda, M. Kojima, C. Kawazu, M. Murata, H. Nishiyori, D. Lazarevic, D. Motti, T. T. Marstrand, M. H. Tang, X. Zhao, A. Krogh, O. Winther, T. Arakawa, J. Kawai, C. Wells, C. Daub, M. Harbers, Y. Hayashizaki, S. Gustincich, A. Sandelin, and P. Carninci. 2009. Genome-wide detection and analysis of hippocampus core promoters using deepcage. Genome Res 19, no. 2: 255-65.

Vazquez-Pianzola, P., G. Hernandez, B. Suter, and R. Rivera-Pomar. 2007. Different modes of translation for hid, grim and sickle mRNAs in drosophila. Cell Death Differ 14, no. 2: 286-95.

Venderova, K., G. Kabbach, E. Abdel-Messih, Y. Zhang, R. J. Parks, Y. Imai, S. Gehrke, J. Ngsee, M. J. Lavoie, R. S. Slack, Y. Rao, Z. Zhang, B. Lu, M. E. Haque, and D. S. Park. 2009. Leucine-rich repeat kinase 2 interacts with parkin, dj-1 and pink-1 in a drosophila melanogaster model of parkinson's disease. Hum Mol Genet 18, no. 22: 4390-404.

Vinson, JP, D DeCaprio, MD Pearson, S Luoma, and JE Galagan. Comparative gene prediction using conditional random fields. Cambridge, MA: The Broad Institute of MIT and Harvard.

Wakaguri, H., R. Yamashita, Y. Suzuki, S. Sugano, and K. Nakai. 2008. Dbtss: Database of transcription start sites, progress report 2008. Nucleic Acids Res 36, no. Database issue: D97-101.

Wang, Y., M. Jorda, P. L. Jones, R. Maleszka, X. Ling, H. M. Robertson, C. A. Mizzen, M. A. Peinado, and G. E. Robinson. 2006. Functional cpg methylation system in a social insect. Science 314, no. 5799: 645-7.

Wang, Z., C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40, no. 7: 897-903.

Wassarman, D. A. and F. Sauer. 2001. Taf(II)250: A transcription toolbox. J Cell Sci 114, no. Pt 16: 2895-902.

Wassarman, L. 2004. All of statistics: A concise course in statistical inference. Edited by G Casella, Fienberg S and I Olkin. Pittsburgh, PA: Springer-Verlag.

Watson, J. D. and F. H. Crick. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, no. 4356: 737-8.

Westwood, J. T. and C. Wu. 1993. Activation of drosophila heat shock factor: Conformational change associated with a monomer-to-trimer transition. Mol Cell Biol 13, no. 6: 3481-6.

Wilhelm, B. T., S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and J. Bahler. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453, no. 7199: 1239-43.

Wilson, R. J., J. L. Goodman, and V. B. Strelets. 2008. Flybase: Integration and improvements to query tools. Nucleic Acids Res 36, no. Database issue: D588-93.

Wolf, M. J., H. Amrein, J. A. Izatt, M. A. Choma, M. C. Reedy, and H. A. Rockman. 2006. Drosophila as a model for the identification of genes causing adult human heart disease. Proc Natl Acad Sci U S A 103, no. 5: 1394-9.

Wright, K. J., M. T. Marr, 2nd, and R. Tjian. 2006. Taf4 nucleates a core subcomplex of tfiid and mediates activated transcription from a tata-less promoter. Proc Natl Acad Sci U S A 103, no. 33: 12347-52.

Yokoyama, K. D., U. Ohler, and G. A. Wray. 2009. Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. Nucleic Acids Res 37, no. 13: e92.

Yuan, G. C., Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. 2005. Genome-scale identification of nucleosome positions in s. Cerevisiae. Science 309, no. 5734: 626-30.

Zavolan, M., S. Kondo, C. Schonbach, J. Adachi, D. A. Hume, Y. Hayashizaki, and T. Gaasterland. 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res 13, no. 6B: 1290-300.

Zeitlinger, J., A. Stark, M. Kellis, J. W. Hong, S. Nechaev, K. Adelman, M. Levine, and R. A. Young. 2007. RNA polymerase stalling at developmental control genes in the drosophila melanogaster embryo. Nat Genet 39, no. 12: 1512-6.

Zhan, M., H. Yamaza, Y. Sun, J. Sinclair, H. Li, and S. Zou. 2007. Temporal and spatial transcriptional profiles of aging in drosophila melanogaster. Genome Res 17, no. 8: 1236-43.

Zhang, M. Q. and T. G. Marr. 1993. A weight array method for splicing signal analysis. Comput Appl Biosci 9, no. 5: 499-509.

Zhang, Y., Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, and K. Struhl. 2009. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol 16, no. 8: 847-52.

Zhang, Z. and F. S. Dietrich. 2005. Identification and characterization of upstream open reading frames (uorf) in the 5' untranslated regions (utr) of genes in saccharomyces cerevisiae. Curr Genet 48, no. 2: 77-87.

_____. 2005. Mapping of transcription start sites in saccharomyces cerevisiae using 5' SAGE. Nucleic Acids Res 33, no. 9: 2838-51.

Zhu, Q. and M. S. Halfon. 2009. Complex organizational structure of the genome revealed by genome-wide analysis of single and alternative promoters in drosophila melanogaster. BMC Genomics 10: 9.

Zofall, M., J. Persinger, S. R. Kassabov, and B. Bartholomew. 2006. Chromatin remodeling by isw2 and swi/snf requires DNA translocation inside the nucleosome. Nat Struct Mol Biol 13, no. 4: 339-46.

Biography

Elizabeth Ann Rach

Born on November 21, 1981 in Buffalo, NY, USA

Education

Duke University	PhD Computational Biology & Bioinformatics (CBB) Thesis: The Spatiotemporal Regulatory Code of Transcription Start Sites	2004 - 2010
Cornell University	B.A. Mathematics Cum Laude Honors Concentration in Biological/Genetic Mathematics Thesis: Power Law Distributions of Gene Family Sizes	2000 - 2004
Phillipsburg High School	Valedictorian	Class of 2000

International Education

Hinxton, England	Wellcome Trust Sanger Center Dros Genomics Course	August 2007
Vienna, Austria	Intelligent Systems for Molecular Biology (ISMB) Conference	July 2007
Dresden, Germany	Max Planck Institute for Molecular Cell Biology & Genetics (MPI-CBG) Independently researched *in situ* hybridization in the fruit fly	June-July 2007
Tuscany, Italy	Cornell University Pauline & Irving Tanner Scholar Independently researched St. Caterina artwork in Firenze, Pisa, Siena, Roma	August 2003
University of Sydney, Australia	Studied music at the Opera House & Conservatory	Feb-June 2002

Honors and Awards

Sigma Xi Research Society	2008-2010*
Duke University Graduate Fellowship	2004-2010*
Cold Spring Harbor Laboratory Travel Award	2008
Duke University Conference Travel Award	2008
Duke University International Research Travel Award	2007-2008
US Department of Energy Travel Award	2007
Burroughs Wellcome Fund & Triangle Community George Hitchings Young Investigator Award	2006-2007
Cornell University Pauline & Irving Tanner Scholarship	2000-2004*
John T. Ward Chemistry Scholarship	2000-2004*
National Society of Collegiate Scholars	2002-2004*
NJ Star-Ledger Scholarship	2000-2004*
Russo and Russo Foundation Scholarship	2000-2004*
George Washington University Award for Excellence in Science & Math	1998

*annual award

Publications

Rach EA, Winter D, Corcoran DL, Ting N, Zhu J, Ohler U. **Nucleosome Organization, Chromatin Structure, and Insulators Indicate Divergent Strategies for Transcription Initiation.** *In review*, March 2010.

Ting N, Corcoran DL, Rach EA, Song S, Spana E, Gao Y, Ohler U, Zhu J. **The Landscape of Transcription Initiation in the *Drosophila melanogaster* Embryo.** *In review Nature Methods,* March 2010.

Rach EA, Yuan HY, Majoros WH, Tomancak P, Ohler U. Motif Composition, **Conservation, and Condition-Specificity of Single and Alternative Transcription Starts Sites in the *Drosophila* Genome.** *Genome Biology,* 2009. 10(7):R73.
Article ranked #6 on Genome Biology's top 20 most accessed articles July 2009.

Malcolm W, Lenfestey R, Rice H, Rach E, Goldberg R, Cotton C. **Dietary Fat for Infants With Enterostomies**. *Journal of Pediatric Surgery*, 2007. 42(11): 1811-5.

Rach E. **Power Law Distributions of Gene Family Sizes.** Cornell University Senior Honors Math Thesis, May 2004.

Presentations

RECOMB Regulatory Genomics Selected Presentation	Broad Institute/MIT	December 2009
Cold Spring Harbor Labs Poster Presentation		March 2008
Duke Institute for Genome Sciences & Policy	Grandover Hotel	September 2007
Wellcome Trust Sanger Center *Drosophila* Genomics Presentation	Hinxton, UK	August 2007
Intelligent Systems for Molecular Biology (ISMB) Poster Presentation	Vienna, Austria	July 2007
Max Planck Institute- Molecular Cell Biology & Genetics Presentation	Dresden, Germany	June 2007
First Place Annual CBB Program Poster Competition		September 2006

Personal

I was born in 1981 to my parents Kathy and Herb Rach, and an older brother Matt. My family lived in Buffalo until I was two, when we moved to Syracuse, NY. In my younger childhood, I enjoyed tap dancing, playing the piano, and playing in the snow! When I was 12, we moved to Stewartsville, NJ, where I played tennis and rode horses (Dance Class) throughout high school. In 2000, I had the great privilege of attending Cornell University in Ithaca, NY, where I studied math and genetics, in addition to riding on the Cornell Equestrian team. Upon graduating, I enrolled in the PhD program for Computational Biology and Bioinformatics at Duke University. For the past 6 years, I have greatly enjoyed reading Christian literature, playing volleyball in the summer medical school league, riding horses (Tex), studying/researching abroad, volunteering in the Duke Hospital Neonatal Intensive Care Unit (NICU), feeding the ducks at my apartment (Ducky and the crew), and eating ice cream. The Lord has blessed me with a fantastic chocolate lab Mochsie, two beautiful Godchildren Jack and Nicole, a sister-in-law named Jenni, and a niece named Julianna, who is expected to be born this May!