March 2021

Genomic Studies of BEAF-32 in Drosophila

John Keller McKowen III
Louisiana State University and Agricultural and Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Part of the Bioinformatics Commons, Genetics Commons, Genomics Commons, and the Molecular Genetics Commons

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/5515

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
GENOMIC STUDIES OF BEAF-32 IN DROSOPHILA

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Biological Sciences

by
John Keller McKowen III
B.S., Louisiana State University, 2012
May 2021
ACKNOWLEDGMENTS

I am very grateful to my Advisor Craig Hart for his guidance and support over the years. As an undergraduate, he gave me opportunities to contribute to various projects. As a graduate student, he allowed me the freedom to explore new ideas and methods. His dedication to finding truth is truly admirable. I am also grateful for all of the opportunities provided to me by LSU. There are so many great people here that have taught me many things. Special thanks to my committee members Dr. Brylinski and Dr. Dassanayake for their advice and support.

I would like to thank my mother and father who were amazing role models for me growing up. Mom, you inspired me to do well in school and fostered my interest in biological sciences. Your love and discipline surely aided me in getting this far. Dad, your hard work and the sacrifices you have made for my benefit and did not go unnoticed. I hope I can be as good a dad as you have been. To my sister Agnes and cousin Garrett, your love and support has been a constant bedrock on which I can rest on.

To my wife, Addie, I love you very much. To me you are perfect. I am very lucky to have been betrothed to such a kind, clever, and love filled person. Life would be hollow if I did not have your love. Thank you for all of you love, support, advice, and constructive criticisms. To my daughter, Clementine, you are my hope and inspiration. Being your dad is wonderful. Your smile and laugh always brighten my day. I look forward to teaching you about nature and life and watching you grow.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS...........................................................................................................ii

ABSTRACT............................................................................................................................iv

CHAPTER 1. LITERATURE REVIEW......................................................................................1

CHAPTER 2. A SEARCH FOR *TOFU*, THE MUTATION COMPENSATING FOR LOSS OF FUNCTIONAL *BEAF*.....................................................................................................................42

CHAPTER 3. GENOMIC INVESTIGATIONS OF BEAF FUNCTION........................................73

CHAPTER 4. CONCLUSIONS................................................................................................121

APPENDIX. SUPPLEMENTARY TABLE FOR CHAPTER 3..................................................128

LIST OF REFERENCES.........................................................................................................129

VITA........................................................................................................................................148
ABSTRACT

Compared to humans, the model organism Drosophila melanogaster is particularly gene dense and encodes several insulator binding proteins (IBP) to aid in its genome organization. Our focus of research is a particular IBP, Boundary Element-Associated Factor of 32kD (BEAF). BEAF primarily binds near the promoters of constitutively active housekeeping genes. It is thought to help maintain the active state of these genes by preventing the spread of repressive chromatin. Additionally, the enhancer blocking activity of BEAF is thought to prevent aberrant activation or silencing of genes. BEAF appears to be ubiquitously expressed among cells and tissues, albeit at varying levels. Despite its common expression, BEAF has an influence in certain developmental pathways, particularly in eye and ovary development. Ovaries from flies lacking BEAF are malformed and produce very few viable eggs. We isolated a mutation that rescues the ovary development phenotype and restores fertility to flies which lack functional BEAF. In this work we search for the location of this mutation we have named Tofu by the use of genomic sequencing, variant analysis, and classical genetics. From our investigations we identified mutations in a silencing element that may cause the misregulation of the ribbon gene. Evidence suggests ribbon may bind similar sequence motifs as BEAF and was identified in a Co-IP using BEAF, making it a strong candidate for Tofu. In a separate effort, we collect and analyze public genomic data relevant to BEAF, and analysis of lab generated data in order to dissect it’s role in insulation and gene activation. This includes: An analysis of BEAF binding in multiple cell and tissue types to look for differential binding and the effects on gene regulation. Analysis of BEAF-associated promoter size and motif distribution. Analysis of effects of transcription by
RNA-seq and PRO-seq after BEAF knockdown. Analysis of nucleosome distribution after BEAF knockdown by MNase-seq. Analysis of chromatin conformation (Hi-C) after BEAF KD. We report that BEAF has a role in activating promoters to which it is bound, and find evidence that it insulates against activation by GAF.
CHAPTER 1. LITERATURE REVIEW

CHROMATIN

Multicellular eukaryotic organisms are composed of many cells, each carrying the same DNA code but displaying vastly different morphologies in different cell and tissue types. In 1941, Avery, MacLeod, and McCarthy published results that revealed genes are contained in deoxyribonucleic acids, DNA (Avery et al., 1944). Before this critical discovery, the observations and experiments of the day showed DNA as rather uniform among species, while proteins held detectable variations. This led to the prevailing paradigm that proteins encoded genes, while DNA was merely a scaffold. After Avery et al.’s study was published, it was clear that DNA determined genetics, but the abundant proteins still fascinated researchers. Accumulating studies began to elucidate the role of proteins in condensing and organizing the DNA polymer. This led to the study of chromatin (Olins and Olins, 2003).

Genes are segments of DNA code which are transcribed into a less stable nucleic acid, RNA. Messenger RNAs containing protein coding genes are processed and transported form the nucleus to the cytosol. Protein coding genes are then translated from the nucleotide sequence by ribosomes (Crick et al., 1961; Matthaei et al., 1962). Protein packaging of DNA regulates its function. Highly condensed DNA (heterochromatin) is less transcriptionally active than decompacted DNA (euchromatin). DNA contains biological information encoded by a sequence of nitrogenous bases adenine (A), thymine (T), cytosine (C), and guanine (G). Regulatory DNA regions such as promoters, enhancers, silencers, and insulators attract a wide array of proteins to influence the transcription of genes. Gene expression changes in response to the environment and
differs in each specialized cell type of multicellular organisms. How these genetic changes are controlled by proteins within cells has been a major focus in the field of Cell Biology.

![Figure 1.1](image.png)

**Figure 1.1.** A structural diagram of single stranded DNA. Used with permission from Britannica eReader.com, a service of Encyclopædia Britannica, Inc.

DNA molecules are polymers of individual nucleotides which are composed of a central 2-deoxyribose sugar, with a nitrogenous base attached to the 1’ carbon and an acidic phosphate group attached to the 5’ carbon. Polymers consist of nucleotides linked together at the 5’ phosphate to the 3’ hydroxyl group of the sugar. This results in a directional 5’ to 3’ polymer. Two DNA polymers with an inverse sequence of nucleotides are arranged anti-parallel so the complementary bases interact and stabilize the molecule. Adenine pairs with thymine, and
cytosine with guanine. The paired polymers have a double helix structure with a complete turn spanning 10.5 nucleotides in the common B form of DNA (Watson and Crick, 1953). This arrangement leads to the phosphate groups being placed on the outside of the DNA polymer, with the paired bases accessible from grooves in the helix. The negative charge of DNA hampers folding and packaging due to electrostatic repulsive forces; therefore, proteins are crucial for DNA compaction (Bloomfield, 1996).

**HISTONES**

A group of proteins called histones contain basic amino acid residues that interact with and stabilize the acidic phosphate backbone of DNA, forming a complex called the nucleosome. Considered the quantum unit of chromatin, the nucleosome is a disk shaped octamer complex of four core histone proteins (H2A, H2B, H3, and H4) onto which 146bp of DNA is wrapped. H3 and H4 form a stable tetrameric complex, while H2A and H2B form two dimers. These subunits interact to form a histone octamer. DNA makes 1.75 turns around the histone octamer, making contact every 10bp (Kornberg and Lorch, 1999). This model of nucleosome structure is called “beads on a string” with 10nm nucleosomal “beads” connected by 2nm linker DNA. Core histones only account for 57% of DNA neutralization. The remaining charge is neutralized by other proteins and positively charged ions (Strick et al., 2001). In heterochromatin, the non-core histone H1 binds linker DNA between nucleosomes, increasing neutralization of chromatin and helping compact the chromatin fiber. When H1 is incorporated into the structure, it is known as the chromatosome (Bharath, 2003). In the classic model, condensed chromatin can be further packaged into more complex structures such as 30nm solenoid fibers (Finch and Klug, 1976). These higher order structures culminate to form mitotic chromosomes. Newer studies, however,
question the 30nm fiber and suggest an alternative “liquid” model where the 10nm nucleosomal chromatin undergoes irregular folding into thicker structures (Maeshima et al., 2010).

Various post translational modifications (PTMs) of histones compact or decompact the chromatin environment, which in turn affects organization and gene regulation. Examples of PTMs include: acetylation and methylation, phosphorylation, ubiquitination, sumoylation, and glycosylation. Modifications are facilitated by enzymes that add or remove functional groups to key amino acid residues, including: histone acetyl transferases HATs, histone methyltransferases HMTs, histone demethylases HDMs, and histone deacetylases HDACs. Generally, acetylation opens the chromatin, and methylation condenses it. As you can see in the figure below, however, the effect of the modification depends on its context [see (Boros, 2012) review]. The chromatin environment is intricate, with models trying to simplify the complexity by dividing the chromatin environment into 5 to 30 chromatin states (Filion et al., 2010; Kharchenko et al., 2011). Each of the various states can be broadly characterized as active or repressive.

Figure 1.2. The effects on nucleosomes from methylation at differing histone sites (From (Sun et al., 2018) Open access figure)
CHROMATIN REMODELING

The position of nucleosomes can be altered by ATP-dependent chromatin remodeling complexes. These complexes affect nucleosomes in three ways: promoting proper spacing and density of nucleosomes; sliding or evicting nucleosomes; and replacing the canonical histones with variants which affect chromatin structure (Clapier et al., 2017). FlyBase, the Drosophila information database, lists ten complexes in this group: ACF, BAP, CHRAC, INO80, NORC, NURD, NURF, PBAP, TIP60 and TORC. These complexes are grouped into four families based upon catalytic subunits: SWI/SNF; ISWI; INO80; CHD.

ISWI type remodeling complexes (ACF, CHRAC, NORC, NURF, and TORC) aid in the maturation, assembly, and proper spacing of nucleosome arrays (Längst and Manelyte, 2015). The SWI/SNF type complexes, BAP and PBAP, increase chromatin accessibility by sliding or evicting nucleosomes (Becker and Workman, 2013; Bouazoune and Brehm, 2006). INO80 family complexes are involved with histone variant exchange in yeast and vertebrates, but in Drosophila, it is similar to ISWI type nucleosome remodeling (Bao and Shen, 2007). TIP60 facilitates the exchange of canonical histone H2A with histone variant H2Av with the aid of histone chaperone complexes and also has HAT activity (Clapier and Cairns, 2009). The NURD complex is a CHD type remodeler that aids in chromatin assembly but also contains a deacetylase subunit with which it participates in incorporating H3.3, a histone variant associated with transcriptionally active chromatin (Murawska and Brehm, 2011).
Regulation of gene transcription

Genes are regulated to specific levels, allowing cells to optimally respond to the environment and differentiate into various forms in multicellular organisms. A simple gene model consists of regulatory regions that control gene activation, a core promoter region at the start site of transcription, and a coding region that contains the functional sequence. Sequence specific Transcription Factors (TFs) bind regulatory regions to control the gene. TFs then recruit chromatin remodeling complexes, histone modifiers, basal transcription factors, and other cofactors to the core promoter. Basal, or general transcription factors (GTFs) recruit RNA Polymerase and are essential for transcription initiation (Ngoc et al., 2019).

Figure 1.3. An illustration of three types of ATP-dependent Chromatin Remodelers. ISWI and CHD are grouped together here. Figure adapted with permission from (Clapier et al., 2017)
TRANSCRIPTION FACTORS

Sequence specific Transcription Factors are a major focal point in the study of gene regulation. These proteins modulate their target sites by recruiting chromatin remodelers, histone modifiers, and general transcription factors. TFs have DNA binding regions that recognize specific or generalized DNA sequences called motifs. Their activity depends on where in the genome they are recruited to and what proteins they interact with. TFs can recruit activating complexes and GTFs to activate genes, or silencing complexes to deactivate genes. TFs can act alone, but will often work synergistically or antagonistically with other TFs to regulate gene function. These interactions can be genetic, where factors facilitate or inhibit one another’s activity, or physical, where they form a complex that affects their respective functions. There are over 700 sequence specific TFs in Drosophila (Hammonds et al., 2013). The expression of each differs among cell types, and the combination of TFs in each cell serves to regulate genes in accordance with the cell’s function.

Additionally, TFs often physically interact with non DNA binding proteins (cofactors) that serve critical functions. Cofactors can either negate or add to the effect of a transcription factor. In this way, they serve a regulatory function that modulates the effects of a TF. Some facilitate the interaction of TFs that do not physically interact by acting as a bridge between the two proteins. By bridging TFs, they help mediate enhancer-promoter interactions. A common but important cofactor is the Mediator complex. Mediator interacts with TFs, GTFs, and RNA pol II. It forms chromatin loops between TF bound enhancers and is essential for RNA Pol II transcription initiation (Soutourina, 2018).
Chromatin-Immunoprecipitation (ChIP) is a technique commonly employed to determine where TFs bind chromatin. ChIP involves raising an antibody specific to a protein of interest. Antibodies can be raised to specifically recognize PTMs such as histone modifications. The chromatin is then cross-linked and fragmented. The antibody is used to immunoprecipitate the protein of interest, bringing with it any cross-linked segments of DNA. The precipitated DNA segments are then purified and mapped to the genome using microarray (ChIP-chip) or high throughput sequencing technology (ChIP-seq). The aligned reads give spatial information for binding sites as well as sequence information that can help determine binding motifs. This technique is also useful for non-TF proteins such as modified histones and RNA-Polymerases (Johnson et al., 2007; Mikkelsen et al., 2007).

RNA POLYMERASE

RNA polymerases synthesize RNA using DNA as a template through a process known as transcription. RNAs have many different functions and different RNA polymerase complexes to synthesize them. RNA Pol I produces most ribosomal RNAs (rRNA) that are used in the ribosome for protein synthesis. RNA Pol III makes many types of smaller RNAs including 5S rRNA, transfer (tRNA), U6 snRNA, 7SL RNA, SINEs, and antisense RNAs (siRNA and miRNA). RNA Pol II transcribes messenger RNA (mRNA) from protein coding genes as well as small nuclear (snRNA) and long non-coding RNAs (lncRNA). Additionally many siRNA and miRNAs are processed out of RNA Pol II transcripts (Sainsbury et al., 2015a). Our studies focus on RNA Pol II transcription.
PROMOTER ASSEMBLY AND INITIATION OF TRANSCRIPTION

RNA polymerases do not bind specific DNA sequences and need to be recruited to promoters. The structure of DNA, as well as chromatin, inhibits the process of transcription. Thus, additional factors are required for RNA polymerase assembly on the promoter, transcription initiation, and elongation. TFs recognize promoters, and recruit chromatin remodelers, histone modifiers, cofactors and general transcription factors (GTFs). GTFs combine with RNA Polymerase to form the pre-initiation complex (PIC). Each polymerase has its own set of GTFs, and each GTF may be a complex of subunit proteins. The most well studied PIC associates with promoters containing a TATA box motif, but most promoters in *Drosophila* lack this motif. First, we will review the formation of the PIC on TATA promoters then discuss variations of the PIC at TATA-less promoters.

The Pol II PIC formation is a multi step process beginning with the recruitment and binding of TFIID which binds the TATA motif with its TATA box-binding subunit (TBP). Then, TFIIA and TFIIB are recruited and bind regions around the TATA motif. RNA Pol II and TFIIF join the complex to form the core PIC. Once the core PIC is formed, TFIIE is able to bind and recruit TFIIH whose helicase activity separates the DNA strands, opening the “transcription bubble.” TFIIH also has kinase activity which phosphorylates Ser5 on the carboxy tail domain (CTD) of RNA Pol II. Phosphorylation alters protein-protein interactions of Pol II and signals the commencement of active transcription (Sainsbury et al., 2015b). Additionally the Mediator complex is needed to promote PIC assembly and initiation. It physically interacts with TFIIB, TFIID, TFIIH, and RNA Pol II and aids PIC assembly at multiple stages. It also helps regulate
the Ser5 CTD phosphorylation of RNA Pol, thus playing a critical role in switching the PIC between poised and active states (Soutourina, 2018).

Many Drosophila promoters do not contain a TATA Box, and in most cases, TFIID contains a TBP-related factor (TRF) instead of TBP. In Drosophila, there are five known TRFs TRF1&3 are interchangeable with TBP, and TRF4&5 are cytoplasmic (Kurshakova et al., 2019). TRF2 is a much larger protein than TBP and the other TRFs, and it does not associate with TFIID. Instead, TRF2 interacts directly with TFIIA and TFIIIB to form an alternative PIC (Andersen et al., 2017; Rabenstein et al., 1999). Though it has homology to TBP, it does not have detectable binding activity (Wang et al., 2014). Instead, it is likely recruited to promoters by TFs such as M1BP and DREF (Baumann and Gilmour, 2017) (Hochheimer et al., 2002).

PAUSING AND ELONGATION

Elongation is the third step of transcription and a highly regulated process. After activation, initiation factors are released, and pausing factors DSIF and NELF bind RNA Pol II in their absence. This causes transcribing RNA Pol II to accumulate at high levels in a region 30-60 bp downstream of the transcription start site (TSS) (Vos et al., 2018). This rate limiting step to elongation serves as a regulatory and quality control checkpoint. Genome wide studies have found that many constitutive genes demonstrate pausing, but developmental and stimulus controlled genes commonly have a high pausing index. The first nucleosome downstream of the TSS (+1 nucleosome) was also shown to act as a barrier for the elongation of RNA Pol II, but it is attenuated by the substitution of the histone variant H2A.Z into the +1 nucleosome (Weber et al., 2014). Therefore it is reasonable to expect TIP60 and similar chromatin remodelers may have a role in pausing.
Release of pausing is mediated by the P-TEFb complex which is recruited to promoters directly or indirectly by TFs and cofactors such as Mediator. P-TEFb functions by phosphorylating the carboxy terminal domain (CTD) of RNA Pol II and pausing factors, NELF and DSIF. After phosphorylation, NELF is evicted and DSIF becomes a positive elongation factor (Jonkers and Lis, 2015; Peterlin and Price, 2006). P-TEFb has been shown to be critical for the expression of almost all active genes (Henriques et al., 2013; Rahl et al., 2010). Although only a subset of active genes are highly paused, these observations suggest that P-TEFb dependent pausing could be an important regulatory step for all active genes (Jonkers and Lis, 2015). After pause release, elongation continues with the aid of histone modifiers, chromatin remodelers, elongation factors, and histone chaperones, all together functioning to displace nucleosomes as Pol II proceeds through the gene (Venkatesh and Workman, 2015). It is thought pausing and release is determined by TFs and co-factors that recruit complexes that control pausing (Jonkers and Lis, 2015).

**CORE PROMOTER**

The PIC interacts with a core promoter region that is positioned around -40 to +40 base pairs around the transcription start site (TSS). A study of core promoters in *Drosophila* found 10 enriched motifs, four of these corresponding to previously found motifs, including: downstream promoter element (DPE), DNA replication-related element (DRE), initiator (Inr), and TATA-box (TATA) (Ohler et al., 2002). There are no universal core promoter motifs, and they are often found in various combinations. Some core promoters have no recognizable motifs. Inr or Inr-like elements are the most commonly occurring sequences that overlap +1 at the TSS. Core promoter analysis is further compounded by the existence of dispersed promoters, resulting in multiple
TSS spread out over a region as large as 100nt. Based on this classification, TATA, Inr, Ohler motif 10 (MTE), and DPE are often found in focused gene promoters. DRE and Ohler motifs 1, 6, and 7, however, are associated with dispersed gene promoters (Hoskins et al., 2011; Ni et al., 2010; Rach et al., 2009). Dispersed promoters tend to be associated with strongly positioned nucleosomal arrays (Rach et al., 2011). Most dispersed promoters correspond with constitutively active genes, while focused promoters tend to be developmentally or environmentally regulated. Exceptions include the promoters of most ribosomal protein genes which are focused yet constitutively active and contain the polypyrimidine initiator (TCT) motif at the TSS (Ngoc et al., 2019). The arrangement of core promoter elements is thought to help position the PIC to select the proper TSS. In TATA-less promoters, TFs such as DREF can directly recruit TRF2 to the core promoter (Hochheimer et al., 2002). There is evidence that TSS selection of TATA-less dispersed promoters is directed by the nonspecific lethal (NSL) complex binding AT-rich sequences in conjunction with other core promoter elements such as DRE (Lam et al., 2019).

Figure 1.4. Schematic diagram of a focused promoter. Figure adapted from (Butler and Kadonaga, 2002) under creative commons license.
PROXIMAL PROMOTER REGION

The proximal promoter region is located directly upstream of the core promoter and contains sequences that attract specific, and often multiple, TFs. The region varies based on TF binding and averages around 300bp in length. Binding this region, TFs can activate genes by recruiting activating chromatin remodelers, co-factors and GTFs to the core promoter to assemble the PIC (Ngoc et al., 2019). Repressor TFs can bind this region and block activation by preventing the binding of activating TFs often recruiting repressive chromatin remodelers. Repressor TFs also play a role in regulating enhancers (Chopra et al., 2012; Mouawad et al., 2020). Thus, the proximal promoter region serves to augment the core promoter, regulating transcription under the control of sequence specific TFs.

ENHANCERS

Enhancers, like proximal promoter regions, are DNA elements that contain sequences specific to particular TFs and help activate genes. Unlike promoter elements, enhancers act at a distance (trans) from the TSS of genes they activate, this distance can range from a thousand base pairs to over a million (Lettice et al., 2002; Riethoven, 2010; Small and Arnosti, 2020). Enhancers interact with promoters to increase transcription levels of genes. A single enhancer is capable of activating multiple promoters, as is the case for enhancers of the engrailed and invected genes in Drosophila, which are separated by over 50kb (Goldsborough and Kornberg, 1994; Kwon et al., 2009). Additionally, multiple enhancers can target a single promoter to (a) provide redundancy, as is the case for Drosophila’s snail and shavenbaby genes (Perry et al., 2010; Preger-Ben Noon et al., 2018; Thurman et al., 2012) or (b) promote a differential expression of the promoter in different cell types, as is the case with the yellow gene (Geyer and
One function of TFs is to bring an enhancer element in physical proximity to a promoter element, allowing the factors bound to the enhancer to influence the target gene by forming loops in the chromatin (Deng et al., 2012; Jing et al., 2008). Some of these enhancer promoter loops have been shown to be associated with paused RNA Pol II (Ghavi-Helm et al., 2014). This evidence suggests enhancers can play multiple roles in gene regulation, rather than simply promoting transcription, or can promote transcription in different ways such as aiding in RNA Pol II recruitment and regulation of pausing and elongation.

Enhancers have been shown to recruit PolII and initiate transcription on their own (Kim et al., 2010; Santa et al., 2010). A common way to detect enhancers is to look for signs of
transcription. A widely held hypothesis is that enhancers evolved from promoter elements (Arenas Mena, 2017). A study found that enhancers and promoter activity can be interchangeable (Mikhaylichenko et al., 2018). Additionally, an unbiased genomic screen of DNA elements, STARR-seq, demonstrated that many promoters can act as enhancers (Arnold et al., 2013).

**SILENCERS**

Silencers, like enhancers, are DNA elements that contain TF binding sites and function irrespective of distance to their target genes. These sites recruit repressor TFs and can block the recruitment of Pol II, prevent binding of activating TFs, or recruit chromatin silencing remodelers (Riethoven, 2010). Polycomb Repressive Elements (PREs) are a common type of silencer that attract Polycomb group proteins, which leads to the spreading of the silencing histone modification H3K27me3. In a recent *Drosophila* study, most of the interrogated silencers were found to act as enhancers in different cell contexts. It was also found that these silencers contain more TF binding sites than a typical enhancer, suggesting silencers are a regulated subclass of enhancers (Gisselbrecht et al., 2020). A study of PREs in the *Drosophila* embryo demonstrated that, in certain contexts, elements classified as developmental enhancers can act as PREs, while the reverse can also be true (Erceg et al., 2017).

**INSULATORS**

Insulators, or boundary elements, are another class of regulatory DNA regions (Modolell et al., 1983; Parkhurst et al., 1988; Udvardy et al., 1985). These sites ensure enhancers contact their target gene promoters and do not instead interact with off-target sites (Hou et al., 2008; Kellum and Schedl, 1992). They have also been demonstrated to act as a barrier to prevent
heterochromatin spreading (Donze et al., 1999; Kellum and Schedl, 1991). Like other regulatory regions, insulators recruit proteins to carry out their functions. These proteins are thus named Insulator Binding Proteins (IBP). Our focus is to study the molecular mechanisms of one such IBP from *D. melanogaster*, Boundary Element Associated Factor of 32Kd (BEAF). Figure 1.5 below illustrates the enhancer blocking role of insulators, it also shows that the ME insulator depends on BEAF to function (Sultana et al., 2011).

Figure 1.6. The diagram on the left shows a schematic representation of enhancer blocking. By removing the boundary element, the enhancer is able to activate the gene for red eyes. On the right you see red eyes when the boundary is removed in the WT, the boundary does not function in the BEAF mutant background so the eyes are red in both cases. The graph shows quantification of eye pigments by spectroscopy. Figure adapted with permission from (Sultana et al., 2011).

Similar to enhancers, insulators are also thought to be evolutionarily related to promoter elements (Raab and Kamakaka, 2010). Promoters with paused polymerase in the *Drosophila* bithorax cluster were shown to have enhancer blocking activity, dependent on the presence of NELF (Chopra et al., 2009). BEAF, while having enhancer blocking activity, is most often found in the promoter region of genes (Jiang et al., 2009a). A recent study from our lab, in which the
BEAF bound scs’ insulator element was dissected, demonstrated that overlapping but distinct sequences direct insulator and promoter activity, further blurring the relationship between the regulatory elements (Maharjan et al., 2020). IBPs are also conceptualized as architectural proteins, helping to fold the chromatin into higher order structures. Recently, methods to elucidate higher order chromatin structure have increased the evidence for this architectural role.

**ENHANCER-PROMOTER SPECIFICITY**

Regulatory models for genes have often been divided into two categories functionally defined by their expression: regulated or developmental genes, and constitutively active or housekeeping genes (Lam et al., 2012; Ulianov et al., 2016). Core promoters from these categories were used to conduct an unbiased genomic screen for elements that enhance the core promoter. The results demonstrated that gene regulation by enhancers can be broadly split into housekeeping and developmental categories. Interestingly, over half of the housekeeping enhancers were found in the promoter region of genes, while developmental enhancers were found distal to the TSS (Zabidi and Stark, 2016).

![Figure 1.7](image)

Figure 1.7. The results of STARR-seq show a separation of housekeeping and developmental enhancers. Figure adapted with permission from (Zabidi and Stark, 2016).
From this discussion, it is clear that the roles of promoters, enhancers, silencers, and insulators are not sharply defined. Instead, the role an element plays depends on its sequence affinity for particular TFs. The opportunity to interact with a TF is based on the amount of each TF available in a particular cell and the chromatin context of the binding site. Additionally, the interactions between TFs further mediate how these elements coordinate gene regulation. A model developed from this concept suggests TFs bring genes together and co-regulate them as a unit or “hub” (Giammartino et al., 2020; Hu and Tee, 2017).

**HIGHER ORDER CHROMATIN STRUCTURE**

So far, we have discussed DNA packaging around histone proteins and histone modifications that can lead to open (active) chromatin or condensed (inactive) chromatin. Chromatin, however, is not arranged linearly. Instead, it is folded in a way that facilitates its compaction in the nucleus without encumbering gene expression.

Evidence of the folding process came from Fluorescent In-Situ Hybridization (FISH) experiments in which whole chromosomes were visualized (Cremer and Cremer, 2010). The experiments showed chromosomes occupy discrete territories in the nucleus. They also showed that the position of chromosomes within the nucleus is non-random and linked to gene regulation. Specifically, gene dense regions are concentrated in the interior, while gene poor regions are found near the periphery (Cremer et al., 2001, 2003; Croft et al., 1999). Another technique used to investigate chromatin folding is Hi-C. The technique originally developed as a nuclear ligation assay (Cullen et al., 1993) which led to chromosome conformation capture (3C) (Dekker, 2002). The chromatin is first cross-linked. Then, a restriction enzyme is used to fragment the DNA. The samples are then diluted to low
concentration to favor intra-molecular ligations over inter-molecular ligation. Finally, ligase is added, and DNA fragments that were brought into proximity by cross-linked proteins become circularized. Circular DNA is then purified and sequenced or PCR amplified to reveal interactions between genomic loci. 3C only detects interactions between two known loci using PCR, but the technique has been modified in subsequent assays. 4C detects interactions between a specified locus and the genome using high-throughput sequencing (Zhao et al., 2006). Hi-C takes the approach to the genomic level by utilizing paired end read sequencing to keep track of interacting DNA fragments (Lieberman-Aiden et al., 2009). ChIA-PET detects genome-wide interactions mediated by a specific protein, essentially combining Hi-C and ChIP-seq (Zhang et al., 2012). To visualize Hi-C data, a matrix is constructed using the interaction frequency between DNA regions and displayed as a heatmap.
The frequency of DNA interactions from Hi-C data, helped develop new models of chromatin arrangement during interphase. First we will focus on mammalian Hi-C studies, then discuss studies from *Drosophila*, noting the differences between the systems. The first publication of Hi-C data suggested there are two major types of chromatin interactions,
arbitrarily named A and B, A compartments were shown to be associated with gene rich, open chromatin, while B compartments are gene poor, closed chromatin (Lieberman-Aiden et al., 2009). Discrete chromosomal regions showing frequent interactions were named topologically associated domains (TADs) (Dixon et al., 2012). TADs are apparent as “pyramids” of frequent interactions in the Hi-C matrix heatmaps. Within TADs, we often see smaller hubs of interaction called “subTADs” (Phillips-Cremins et al., 2013). In single-cell Hi-C studies, long distance interactions were variable among cells but when pooled, they resemble TADs seen in other studies. Thus, single-cell Hi-C shows TADs and subTADs are the result of dynamic interaction probabilities rather than firm domains (Flyamer et al., 2017; Stevens et al., 2017). Larger TADs are associated with inactive chromatin (B-compartment), while smaller TADS and inter-TAD regions tend to be gene-dense and active (A-compartment) (Hou et al., 2012). It was found that the insulator binding factor CTCF is present at TAD borders, and deletions of these CTCF-bound borders can disrupt TADs and can cause misregulation of genes (Franke et al., 2016; Lupiáñez et al., 2015). In addition to TADs, long distance interactions, guided by proteins creating chromatin loops, appear as spots on the Hi-C heatmap (Rao et al., 2014). Often, a strong interaction spot will be apparent at the top of a TAD pyramid, and this is thought to result from proteins at the two TAD borders interacting together to create a “loop domain”. In mammals, loop domains are dependent on the cohesin complex (Alipour and Marko, 2012; Rao et al., 2017; Sanborn et al., 2015). In this loop extrusion model, cohesin randomly associates with DNA, bidirectionally extruding it through its ring structure and creating a loop. The looping ceases upon encountering a bound CTCF site, the other side continues looping until blocked again by another CTCF site. CTCF binds DNA asymmetrically, and the CTCF sites in a cohesin-mediated loop must be in a
convergent orientation (Fudenberg et al., 2016). Enhancer-promoter loops are also apparent in Hi-C data as spots, but are not always contained within a TAD and do not define TAD borders (Lu et al., 2020). Much of this work was done in mammals where CTCF is the major IBP, but *Drosophila* TAD formation has notable differences.

In *Drosophila*, loops between TAD borders are seldom seen, and cohesin-CTCF loops are not found in *Drosophila* (Rowley et al., 2017). Instead, TAD formation is linked to active and inactive transcription (A and B compartments), but IBPs also seem to be involved. In mammals, CTCF is the main IBP, but in *Drosophila*, there are several IBPs. Many of these (such as dCTCF, CP190, M1BP, Pita, and BEAF) are found at TAD boundaries and may play roles similar to CTCF in TAD formation and stability (Ali et al., 2016; Cubeñas-Potts et al., 2017; Ramírez et al., 2018; Rowley et al., 2017). Cohesin and the related protein, Condensin II, were also found at *Drosophila* TAD borders. Cohesin colocalizes with CP190, a cofactor of many IBPs, including BEAF and dCTCF. (Bartkuhn et al., 2009; Van Bortle et al., 2014). Furthermore, cohesin was shown to be critical for certain enhancer-promoter interactions and the formation of gene loops between TSS and transcription termination sites (Khaminets et al., 2020; Rowley et al., 2019). However, interaction loops between TAD borders as seen in mammals is uncommon in *Drosophila* (Szabo et al., 2019). Finally long-distance loops in *Drosophila* are most often associated with repressive Polycomb group proteins. These loop anchors are enriched with cohesin but not CTCF (Eagen et al., 2017).
Figure 1.9. A model for genome organization based on HiC and FISH observations. (A) Chromatin loops, including promoter-enhancer loops, form through the close physical proximity of genome regions. Their interactions are stabilized by the formation of phase-separated aggregates (shaded in red) of relevant chromatin proteins and transcription factors (blue squares). Enhancers (yellow nucleosomes) associate with gene promoters (green nucleosomes) to activate gene targets (green arrow) either in single-enhancer loops (left) or in superenhancers (right). (B) TADs form via loop extrusion generated by the cohesin motor (light blue). The CTCF protein (purple) defines the boundary of the domain (gold) by determining the location of cohesin. The internal structure of TADs is determined by heterogenous polymer chromatin-chromatin self-interactions, which are likely stabilized by phase separation (shaded in red). Black arrows denote the direction of movement of DNA through the cohesin complex. (C) Chromatin compartments (left) form via the association of multiple homotypic domains (1–6; green and red) to ultimately form chromosomes (middle). Multiple chromosomes associate via interactions between homotypic domains across the 3D structure of the nucleus to form large-scale blocks of heterochromatin (red) and euchromatin (green) (right). Nuclear bodies (blue) serve as anchoring points. Transcription hubs (yellow) form via the coalescence of multiple active genome regions located on multiple chromosomes. Figure reproduced with permission from (Misteli, 2020).
Another model often used to explain higher order chromatin structure is the liquid-liquid phase separation (LLPS) model (Hnisz et al., 2017). The foundations of this model come from observations of “membraneless organelles” found in the nucleus. This includes nuclear speckles, Cajal bodies, and the nucleolus, which all have liquid-like properties (Mao et al., 2011; Zhu and Brangwynne, 2015). Biochemically, these phases arise from multivalent protein-protein interactions. Proteins that participate in LLPS often have intrinsically disordered regions (IDR) which are thought to facilitate multivalent interactions. The chemical forces that drive phase separation ultimately result from amino acid residues in the IDRs and include ionic, hydrogen bonding, planar, and hydrophobic interactions (Dignon et al., 2020). Phase separation can be regulated by PTMs, especially phosphorylation, which rapidly causes self assembly into phase droplets or dissolution back into the aqueous nucleoplasm (Palikyras and Papantonis). RNA Pol II seems to participate in phase switching through phosphorylation of its CTD during recruitment, initiation, and elongation, suggesting phase separation plays a role in active transcription hubs (Boehning et al., 2018; Kwon et al., 2013; Lu et al., 2018). A heterochromatin protein found in humans, HP1α, was also found to participate in phosphorylation driven phase separation, suggesting heterochromatin may be segregated by LLPS interactions (Larson et al., 2017). These observations suggest LLPS helps drive chromatin organization, and it can be regulated by PTMs that result in phase switching.

**BEAF**

Polytene chromosomes of *Drosophila* salivary glands have undergone 10 rounds of DNA replication with no cell division, resulting in 1024 paired sister chromatids. Polyteny provides an excellent model with which to study genetics and chromatin. Visually, compact heterochromatic
DNA appears as dark bands, while uncompacted euchromatin appears as light inter-band regions. This banding pattern gave early geneticists a physical reference and allowed them to map the location of genes to cytological inter-bands on the chromosome. Interest in this model system led others to observe disruption of inter-bands called “puffs” at (a) certain developmental time points and (b) in response to environmental cues such as heat shock. These puffs are the result of a high level of transcription contained to a specific locus (Zhimulev and Koryakov, 2009; Zykova et al., 2018). Models of how the puffs are constrained to particular loci led to the proposal of boundary elements, now called insulators, which serve to contain the puff. A study of the 87A7 heat shock puff identified two boundaries surrounding the puff named special chromatin structures, scs and scs’ (Udvardy et al., 1985). A protein that binds scs’ was purified and named boundary element-associated factor of 32 kDa (BEAF). BEAF was shown to immunolocalize to many inter-band and puff regions of polytene chromosomes and is present on mitotic chromosomes. DNA footprinting of BEAF on scs’ identified the CGATA motif as its primary target sequence (Zhao et al., 1995).

Two isoforms of BEAF exist: 32A and 32B. They differ only in their N-terminal DNA binding regions, BED zinc finger domains. These isoforms appear to form dimers, trimers and possibly larger complexes. Immunolocalization showed 32A and 32B bind can together or independently on the polytene chromosome and that the relative ratios of the isoforms vary among loci (Hart et al., 1997). It was found that only 32B, not 32A is needed for normal growth and development (Roy et al., 2007a) The C-terminal end of the common region of BEAF isoforms contain a BESS domain needed for self-interactions (Avva and Hart, 2016). The N-
terminal of the common region region has no predicted structure, but was was shown to mediate interactions with Sry-delta (Dong et al., 2020).

BEAF undergoes post translational modification. Early studies observed a phosphorylated portion of BEAF from nuclear extracts (Hart et al., 1997; Zhao et al., 1995). A later study found evidence for glycosylation and phosphorylation, with evidence suggesting that are six forms of PTM BEAF. This study predicted target sites for various PTMs (Pathak et al., 2007). A more recent study linked glycosylation of BEAF with the active chromatin mark H3K4me3 (De et al., 2017). Many of the predicted PTMs are located in the unstructured middle region of the protein. Recent LLPS models, lead us to speculate that these modifications of BEAF may mediate phase-separated regulation.

It was demonstrated that BEAF binding sites are involved in boundary activity, helping shield a reporter gene from the local chromatin environment (Cuvier et al., 1998; Gilbert et al., 2006; Roy et al., 2007a). It was later shown that placing a BEAF-associated insulator site between an enhancer and promoter can block communication between the two elements, known as enhancer blocking activity, and this activity is dependent on BEAF (Gilbert et al., 2006; Roy et al., 2007a; Sultana et al., 2011). These studies show that insulators bound by BEAF are able to aid in the separation of differentially regulated genes and prevent aberrant regulation.

A study using knockout alleles of BEAF further characterized its function. BEAF expression is not essential to survival from embryo to adult, as long as it is provided to embryos maternally. It does, however, play a critical role in ovary development, and BEAF-KO females are nearly infertile. Loss of BEAF causes slight defects in the ommatidia of the eye, resulting in a rough eye phenotype. Loss of BEAF also causes a disruption in the structure of the male X polytene
chromosomes (Roy et al., 2007a). These phenotypes are greatly enhanced by the introduction of a dominant negative allele of BEAF (BID) that lacks the DNA binding domain. This BID transgene inhibits the DNA binding of BEAF complexes. The more exaggerated phenotypes observed in BID flies indicate BEAF may be interacting with other TFs in complex interactions that require BEAF DNA binding. BID also disrupts any maternally provided BEAF from functioning and causes embryonic lethality. (Gilbert et al., 2006). These studies also demonstrate that BEAF is required for the insulator activity of scs’ through enhancer-blocking and position-effect assays.

ChIP studies for BEAF have been performed by a number of labs. Our lab identified 1818 genomic sites based on peak reproducibility among several embryo ChIP-chip samples using three antibodies, one specific for each isoform and one recognizing the region common to both isoforms. Analysis of our labs conservative loci shows 85% of BEAF binding sites are localized near the TSS. The majority of these genes are actively transcribed in most cell types, so they can be classified as housekeeping genes (Emberly et al., 2008; Jiang et al., 2009b; Lam et al., 2012). Motif analysis of ChIP sites confirms CGATA as a consensus binding motif of 32B. One motif, however, is not enough to ensure BEAF binding and there are many CGATA motifs in the genome not bound by BEAF. Moreover, 7.6% of the sites BEAF reproducibly binds contain no CGATA motif at all (Jiang et al., 2009b). This suggests DNA binding may be contextually facilitated with the aid of other proteins at certain sites.

BEAF is expressed ubiquitously (Contrino et al., 2012). Knockdown experiments have shown that BEAF typically serves to keep associated genes active, though a repressive effect was sometimes seen (Jiang et al., 2009a; Lhoumaud et al., 2014). Data from a luciferase reporter
assay demonstrates BEAF is capable of activating promoters (Dong et al., 2020). However, the viability of flies lacking a functional copy of BEAF suggests that it’s activity is redundant in most contexts.

A rough eye assay showed that several developmental TFs have genetic interactions with BEAF (Roy et al., 2007b). Other labs have found genetic interactions with developmental pathways such as BMP (Deignan et al., 2016), and Hippo (Jukam et al., 2016). Multiple lines of evidence suggest (a) a physical interaction between BEAF and Sry-delta and (b) that this interaction can mediate enhancer-promoter communication (Dong et al., 2020). Interactions with developmental transcription factors, as well as eye and ovary developmental defects, suggest BEAF plays a role in developmental regulation.

**RECENT AND ONGOING INVESTIGATION**

In the following chapters, I describe the published and unpublished results of my research into BEAF, with a focus on genomics data. First, I present work investigating Tofu, a mutation that compensates for the loss of BEAF, restoring fertility to females. Then, I present a set of bioinformatic analyses of BEAF related genomics data: an analysis of differential BEAF binding between cell types using ChIP-seq and RNA-seq, an analysis of core promoter and other motifs in relation to promoter architecture and BEAF binding, our work characterizing the global effects of BEAF depletion on transcription by RNA Pol II (PRO-seq) and nucleosome distribution (MNase-seq), and finally the effects of BEAF depletion on genome organization from Hi-C data. The results show a relationship between BEAF binding and transcriptional activation. We also find evidence that BEAF blocks activation by GAGA transcription factor (GAF).


39


CHAPTER 2. A SEARCH FOR TOFU,
THE MUTATION COMPENSATING FOR LOSS OF FUNCTIONAL BEAF

INTRODUCTION

BEAF is an insulator binding protein of *Drosophila melanogaster* that mainly binds the promoter region of housekeeping genes (Cuvier et al., 1998; Jiang et al., 2009; Zhao et al., 1995). The *BEAF* gene is located on the second chromosome. It is composed of three exons and expressed as two isoforms: 32A and 32B. Each isoform has a unique DNA binding region coded by exons 1 and 2, respectively. The third exon contains interacting domains and is shared by both isoforms (Hart et al., 1997). Previously, our lab generated and characterized a nonfunctional allele of *BEAF* by (a) eliminating the ATG start codon of both the 32A and 32B specific exons and (b) introducing two tandem stop codons into the shared exon, *BEAF*\(^{AB-KO}\) (*ABKO*). Flies homozygous for the *ABKO* allele are sickly, display a mild rough eye phenotype, and have defects in oogenesis. Ovaries of *ABKO* flies have malformed egg chambers, displaying variable phenotypes that are most apparent at stage eight of oogenesis. The outcome of these defects is that very few mature oocytes are produced. This phenotype leads to a sharp loss in the fecundity of this fly line. Male fertility is not noticeably affected in homozygous *ABKO* flies. (Roy et al., 2007).

A separate allele, *BEAF\(^{NP6377}\) (NP6377)*, was generated by a randomly inserted Gal4 enhancer trap element containing a mini-white marker gene, *P(GawB)*, into the shared exon of the 32A and 32B isoforms (Hayashi et al., 2002). This *NP6377* line was used in a study that attributed the striking phenotypes of neoplastic growth and recessive larval lethality to the loss of BEAF (Gurudatta et al., 2012), which was not observed in our mutant lines. Skeptical of these
findings, we obtained the \textit{NP6377} line to test for second site mutations. We attempted to rescue the phenotypes using a trans-gene of \textit{BEAF}. If the phenotypes are caused by non-functional \textit{BEAF}, a \textit{BEAF} trans-gene should rescue back to wild type. \textit{ABKO} allele phenotypes were rescued by a trans-gene, but \textit{NP6377} was still recessive larval lethal. The lines were crossed in a complementation test. Since they were independently generated second site mutations would be heterozygous in the progeny. The test showed a rescue of the recessive lethal and neoplastic growth phenotypes, suggesting they were not a consequence of the loss of \textit{BEAF}. \textit{NP6377/ABKO} flies were viable. The defect in fertility caused by the lack of \textit{BEAF} was apparently suppressed by a dominant mutation. \textit{ABKO/ABKO (Tofu KO)} and \textit{NP6377/NP6377 (Tofu NP)} lines were created through meiotic recombination of the \textit{ABKO} and \textit{NP6377} lines. The recombinant lines lacked functional \textit{BEAF} and the recessive lethal mutations, but they retained the dominant enhancer of fertility mutation (Figure 2.1) (Hart, 2014). We named the mutation \textit{Tofu}, because it allows flies to live without \textit{BEAF}. In an attempt to identify \textit{Tofu}, we employed two methods: high-throughput genomic DNA sequencing and classical genetic mapping via meiotic recombination.
Figure 2.1. A: Line drawing of typical fly ovary morphology. Ovaries are composed of several ovarioles which produce oocytes. B: Oocyte development proceeds from the stem cells in the germarium as egg chambers which can be classified as stages by phenotype. C: Ovary phenotypes of $\gamma$-w-, ABKO, and Tofu KO flies. Tofu KO ovaries have mature eggs, but are atypical. (A&B reproduced with permission from (Andersen and Horne-Badovinac, 2016)
METHODS

Fly Stocks

Flies were maintained on standard cornmeal, yeast, and sugar medium with Tegosept. \textit{BEAF}^{NP6377}/CyO \textit{GFP} fly stocks were kindly provided by Victor Corces. \textit{BEAF}^{AB-KO}/G30 fly stocks were previously generated by our lab (Roy et al., 2007b). \textit{BEAF}^{Tofu \textit{NP}} and \textit{BEAF}^{Tofu \textit{KO}} were created by meiotic recombination of \textit{BEAF}^{AB-KO} and \textit{BEAF}^{NP6377} chromosomes (Hart, 2014). Baylor P mapping kit lines, \textit{Actin 5C-Gal4}, \textit{UAS-dpp}, and balancer chromosome lines were obtained from the Bloomington Stock Center.

Fly Genotyping

To detect the \textit{ABKO} allele, we designed a genotyping primer set consisting of three primers. The 3’ BEAF-stop+89-3 primer was common to each set. The BEAF-mut-5 primer has 3’ homology with the introduced two tandem stop codon of the \textit{ABKO} allele. The BEAF-wt-Bam-5 has 3’ homology to the WT \textit{BEAF} allele at the same location as the BEAT-mut-5 primer. The mut-5 and the stop-3 primers give a 536bp primer only from the ABKO allele. The wt-Bam-5 and stop-3 give a 533bp primer only from WT allele. Each primer set was used to genotype flies. Genotyping was done using PHIRE tissue direct master mix.

BEAF-wt-Bam-5: GACATCATATACAGCGAGGA TCC

BEAF-mut-Bam-5: AAGGACATCATATACAGCGAGTAATG

BEAF-stop+89-3: TTACGACACGCTGATTTGCC

DNA Preparation for Sequencing

50 third instar wandering larvae from \textit{y-w-}, \textit{ABKO/CyO-GFP}, \textit{NP6377/CyO-GFP}, \textit{Tofu KO}, \textit{and Tofu NP} lines were harvested. Larvae homozygous for \textit{ABKO} and \textit{NP6377} were screened by
lack of GFP expression from the *ABKO/CyO-GFP*, and *NP6677/CyO-GFP* lines. Larvae were homogenized in 500 µL of Buffer A (10 mM Tris-Cl (pH 7.5), 60 mM NaCl, 10 mM EDTA, 150 uM spermine, 150uM spermidine, 200 ug/mL Proteinase K). Then 500 µL of Buffer B (200 mM Tris-Cl (pH 7.5), 30 mM EDTA, 2% SDS, 200 ug/mL Proteinase K) was added, and the samples were incubated for one hour at 37 C. Samples were then purified by phenol extraction, followed by phenol-chloroform-isooamyl alcohol (25:24:1) extraction, and then chloroform extraction. Samples were finally cleaned by ethanol precipitation and dissolved in 10 mM Tris.

**Illumina Sequencing**

Genomic DNA was submitted to the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana. Libraries were prepared for each sample. Paired end sequencing was done using an Illumina HiSeq2500 instrument.

**Data Analysis**

Raw reads were obtained as paired fastq files. Read quality was analyzed using FASTQC (Andrews, 2010) (Andrews, 2010). Reads were aligned to the dmel r6.09 genome using bowtie2. Samtools was used to assess alignment quality and compress sam alignments into bam format. Variant calling was done using the GATK toolkit (Auwera et al., 2013; DePristo et al., 2011; McKenna et al., 2010). Variant call files (VCF) were compared using bedtools (Quinlan, 2014). Protein coding changes were called with SnpEff (Cingolani et al., 2012). SNP density was visualized as bigwig files using 2000bp windows. Data for recombination rates in 2R was kindly provided by Josep Comeron in 50Kb intervals along the chromosome (Comeron et al., 2012). This data was converted to bigwig format for visualization on IGV. File conversion was done...
using unix commands, kent utilities (Kent et al., 2010), bedTools, and deepTools (Ramírez et al., 2016). Data was displayed using IGV (Robinson et al., 2011).

**Tissue Culturing**

Ovaries were hand dissected from 30 4-day old flies under sterile conditions for each attempt. Following a protocol from (Niki et al., 2006). The egg chambers were disrupted and the tissue was plated in 96-well tissue culture plates. Ovarian tissue from each line was plated in Schneider media supplemented with 5% FBS, 5% Fly extract, and antibiotics (100 U/ml penicillin, 0.1 mg/ml streptomycin). The extract was made from 4-day old flies heterozygous for a *UAS-dpp* trans-gene driven by *Actin 5C-Gal4*. *Dpp* is a growth factor that inhibits differentiation of stem cells, facilitating cell culture (Chen and McKearin, 2003). Once confluent, the cells were transferred to larger culture vessels.
RESULTS

Identification of *Tofu* Chromosome

Casual observations indicated that *Tofu* is on the same chromosome as *BEAF*. We performed a segregation analysis to formally identify the chromosome that contains *Tofu*. This is done by crossing the line carrying the mutation with balancer chromosomes and tracking the phenotypes of the mutant allele along with marker alleles of the balancers. A balancer chromosome has undergone inversions to such an extent that it will not homologously pair with a wild type (WT) chromosome and so suppresses meiotic recombination. They typically contain dominant allele markers and recessive lethal alleles. *CyO* is a balancer of chromosome 2 and contains a dominant marker mutation that gives flies a curly wing phenotype. *TM3* is a balancer of chromosome 3 and contains a dominant marker mutation (*Sb[1]*) that gives flies a “stubble” bristle phenotype. Male flies (XY) can be used to track mutations on the X chromosome. The *ABKO* mutation does not affect male fertility, only females. The *ABKO* fly line is maintained using a *CyO GFP* chromosome. Maternally provided *BEAF* from the *CyO GFP* chromosome is sufficient to produce adult flies homozygous for the *ABKO* allele, zygotically transcribed *BEAF* is not needed for survival of the flies. We performed a series of segregation crosses to determine the chromosome on which *Tofu* resides.
In the first cross (Figure 2.2), we mated Tofu KO flies with a double balancer line CyO/Sp1;TM3/ET50. Virgin female progeny from this cross were mated to homozygous ABKO males. Females from this cross with the TM3 chromosome were isolated with y-w- males and scored for fertility. If Tofu is on chromosome 3, it would become segregated from TM3 and the flies would be infertile. If it is on chromosome 2, the flies would be fertile. If it is on chromosome X or 4, fertility would be around 50%. We assayed 200 flies from the cross and found a fertility rate of 53%. This indicates Tofu is not on chromosome 3, but it does not strongly support association with chromosome 2, unless Tofu, a dominant mutation, was not homozygous in the first cross. We produced flies homozygous for Tofu using the CyO balancer chromosome, indicating Tofu is not recessive lethal. There are no balancer chromosome for chromosome 4 due to its small size, making segregation analysis difficult. Furthermore chromosome 4 contains very few genes, we therefore excluded it from further segregation analysis.

Figure 2.2. Diagram of a segregation cross for chromosomes 2 and 3. An identical mating strategy is divided into models of expected results.
In a second cross (Figure 2.3), we mated males of the Tofu KO line to ABKO/CyO virgin females. The chromosome X will be segregated from male progeny of this cross because male flies receive the maternal X chromosome. Male progeny from this cross were again mated with ABKO/CyO virgin females. If Tofu is on chromosome X, it would have been segregated, and the female progeny would be infertile. If it is on chromosome 2, 50% of female progeny would be fertile. We assayed 200 female progeny from the second cross and found a fertility rate of 52%. This indicates Tofu is not on chromosome X and suggests it is associated with chromosome 2.

Genomic Sequencing

To find the Tofu mutation, we used genomic sequencing to compare mutations among fly lines. We collected genomic DNA from five lines: y-w-, ABKO, NP6377, and the two lines generated from recombination: Tofu NP and Tofu KO. NP6377, Tofu NP, and Tofu KO have the
Tofu mutation, while y-w- and ABKO do not. We used y-w- as a pseudo-wildtype because it is the genetic background the ABKO flies were created in. We planned to identify variants from the reference genome for each line and cross compare variants according to the Tofu phenotype in order to find causal mutation candidates.

We sequenced the genomes using Illumina to generate paired end reads of about 150bp at a genome coverage of over 40 reads per base. In this process, the DNA is fragmented into short segments around 200-800bp in length. Bar-coded forward and reverse adapter sequences are ligated to the fragments, and the reverse adapter has a sequence attached to the flow cell. The fragment is sequenced from the forward primer, and each base is detected and recorded as it is synthesized. Then, the process is repeated on the opposite strand with the reverse primer. This results in high-confidence base calling, which is needed for variant detection (Bentley et al., 2008). After quality checking and adapter trimming, we aligned the reads to the dm6 genome release for D. melanogaster using bowtie2. The alignment process matches the short reads to a pre-computed reference genome. Reads belonging to a pair are coordinated. Paired read sequencing ensures high accuracy, and a non-mapping sequence between read pairs can indicate an insertion mutation relative to the reference sequence (Langmead and Salzberg, 2012; Li et al., 2009).

Table 2.1. Genomic alignment statistics

<table>
<thead>
<tr>
<th></th>
<th>Raw Reads</th>
<th>Paired &amp; Mapped</th>
<th>Mean quality (PHRED)</th>
<th>Mean read size (bp)</th>
<th>Mean fragment size (bp)</th>
<th>Mean coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>y-w-</td>
<td>16188061</td>
<td>16028674</td>
<td>36</td>
<td>130</td>
<td>355</td>
<td>40</td>
</tr>
<tr>
<td>NP63</td>
<td>19202205</td>
<td>19035482</td>
<td>36</td>
<td>130</td>
<td>361</td>
<td>47</td>
</tr>
<tr>
<td>ABKO</td>
<td>17070642</td>
<td>16853622</td>
<td>36</td>
<td>129</td>
<td>363</td>
<td>42</td>
</tr>
<tr>
<td>NPNP</td>
<td>16186293</td>
<td>16010932</td>
<td>36</td>
<td>129</td>
<td>372</td>
<td>40</td>
</tr>
<tr>
<td>KOKO</td>
<td>19677399</td>
<td>19177162</td>
<td>33</td>
<td>150</td>
<td>347</td>
<td>48</td>
</tr>
</tbody>
</table>
After alignment, we called variants by comparing each experimental sequence to the reference genome. To do this, we used a suite of programs in the genome analysis toolkit GATK. Using GATK, we generated variant call files (VCFs) which detail the location of every variation from the reference genome. Changes to a single base are single nucleotide polymorphisms (SNP). Extra bases (insertions) and missing bases (deletions) are lumped together as INDELs. For simplicity of discussion, INDELs are included with SNP variants and together are referred to as SNPs, though they differ biologically. Each line examined had over 700,000 SNPs from the reference genome. After cross comparing, we found 62,045 SNPs that fit the pattern for Tofu among the lines sequenced. Of these, 1,203 were localized to chromosome 2, on which we determined the mutation must reside by the use of balancer chromosomes. The number of SNPs that fit the Tofu pattern on chromosome 2 was small compared to other chromosomes. We reasoned this was due to artificial selection from our crosses and natural selection due to fitness of the Tofu mutation on that chromosome.

Table 2.2. Top, SNP counts per chromosome in sequenced lines. Bottom, SNPs after filtering steps. Files were compared in the order shown.

<table>
<thead>
<tr>
<th></th>
<th>chr2L</th>
<th>chr2R</th>
<th>chr3L</th>
<th>chr3R</th>
<th>chrX</th>
<th>chr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>size (Mb)</td>
<td>23.51</td>
<td>25.29</td>
<td>28.11</td>
<td>32.08</td>
<td>23.54</td>
<td>1.35</td>
</tr>
<tr>
<td>NP6377</td>
<td>164336</td>
<td>148241</td>
<td>181660</td>
<td>149139</td>
<td>121175</td>
<td>2160</td>
</tr>
<tr>
<td>Tofu_NP</td>
<td>215088</td>
<td>192165</td>
<td>192113</td>
<td>179545</td>
<td>113949</td>
<td>1873</td>
</tr>
<tr>
<td>Tofu_KO</td>
<td>168157</td>
<td>145574</td>
<td>191077</td>
<td>179090</td>
<td>106303</td>
<td>1374</td>
</tr>
<tr>
<td>ABKO</td>
<td>166806</td>
<td>145118</td>
<td>229149</td>
<td>186379</td>
<td>119325</td>
<td>1401</td>
</tr>
<tr>
<td>y-w-</td>
<td>167008</td>
<td>148869</td>
<td>209960</td>
<td>150515</td>
<td>104657</td>
<td>1157</td>
</tr>
<tr>
<td>filter_1</td>
<td>133718</td>
<td>141739</td>
<td>155591</td>
<td>118023</td>
<td>73865</td>
<td>1735</td>
</tr>
<tr>
<td>filter_2</td>
<td>85073</td>
<td>79214</td>
<td>150063</td>
<td>106045</td>
<td>71587</td>
<td>901</td>
</tr>
<tr>
<td>filter_3</td>
<td>137</td>
<td>1744</td>
<td>39091</td>
<td>23420</td>
<td>10202</td>
<td>62</td>
</tr>
<tr>
<td>Tofu Pattern</td>
<td>72</td>
<td>1130</td>
<td>34716</td>
<td>20946</td>
<td>4774</td>
<td>57</td>
</tr>
</tbody>
</table>
We used SnpEff to predict the effect of the SNPs on protein synthesis and found 587 of the SNPs on chromosome 2 would cause miss-sense or frame-shift mutations. This number was still far to high to test each mutation in transgenic flies. However we could select candidates from this list to test based on gene ontology. It is also possible that Tofu, a dominant mutation, could be a regulatory mutation, which are more difficult to predict.

<table>
<thead>
<tr>
<th>DNA sequence</th>
<th>Amino acid sequence</th>
<th>Type of mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATG CAG GTG ACC TCA GTG</td>
<td>M Q V T S V</td>
<td>None</td>
</tr>
<tr>
<td>ATG CAG GTT ACC TCA GTG</td>
<td>M Q V T S V</td>
<td>Silent</td>
</tr>
<tr>
<td>ATG CAG TTG ACC TCA GTG</td>
<td>M Q L T S V</td>
<td>Conservative</td>
</tr>
<tr>
<td>ATG CCG GTG ACC TCA GTG</td>
<td>M P V T S V</td>
<td>Missense</td>
</tr>
<tr>
<td>ATG CAG GTG ACC TGA GTG</td>
<td>M Q V T STOP</td>
<td>Nonsense</td>
</tr>
<tr>
<td>ATG CAG GTG AAC CTC AGT G</td>
<td>M Q V N L S</td>
<td>Frame-shift</td>
</tr>
</tbody>
</table>

### Mapping by Meiotic Recombination

To narrow down a region that contains Tofu and thus limit our candidate SNPs, we tried a classic fly genetics technique; the meiotic mapping or genetic linkage assay. The assay takes advantage of homologous recombination of the chromosomes during meiosis. This assay is improved by the fact that meiotic homologous recombination is prevented in male flies, occurring only during oogenesis. Thus, we can limit any crossing over event to a single female fly. During recombination, matching regions in the chromosome are randomly exchanged. The closer together genes are located along the chromosome, the more likely they are to remain together after the exchange takes place. By tracking phenotypes of these genes in the resulting generation, one can calculate a recombination frequency between the genes. If the recombination frequency is less than 50%, the genes are close enough together to be considered linked. Smaller
recombination frequencies indicate closer distance (Bridges, 1935; Morgan, 1911; Sturtevant, 1913).

Calculating recombination rates depends on tracking an observable phenotype. Tofu is a dominant mutation that restores fertility in the absence of functional BEAF. The ABKO and Tofu KO lines have a non-functional allele of the white gene, giving them white eyes (w-). For mapping, we used fly lines from the Baylor P mapping kit. These are w- lines with a white transgene P[w+] inserted into the 2nd chromosome at a known cytological location (Zhai et al., 2003). To track the Tofu phenotype, we first had to replace the wild type BEAF allele of the P[w+] lines with ABKO so that the P[w+] lines would have nonfunctional BEAF (Figure 2.4). To produce the P[w+] KO lines (red eyes), homozygous ABKO males were mated to homozygous P[w+] virgin female flies (P). All (F1) progeny from the (P) cross were heterozygotes, and meiotic recombination of the P[w+] and ABKO alleles may occur in the female’s oocytes. Female (F1) heterozygotes were crossed to male CyO/Sp1 for the recombination cross. Red-eyed females were isolated in the pupal stage and allowed to eclose. Next individual P[w+]/CyO (F2) females were crossed to CyO/Sp1 so all P[w+] progeny have the same chromosome. P[w+]/CyO progeny (F3) were self-crossed. In the next generation (F4) these lines were scored for the ABKO allele using PCR with primer pairs specific for WT or ABKO alleles of BEAF. Once recombination was confirmed, the P[w+] KO/CyO populations were kept as a stable line and named according to their cytological location (e.g. 27E6).
Figure 2.4. Strategy to combine ABKO allele onto P[w+ lines. Recombination occurs in the F1 generation females and will be passed to F2 progeny. Chromosomes are then isolated in the F3 self-cross and the isolated F4 lines were genotyped using primer pairs specific for the WT or ABKO alleles of BEAF. An example of the PCR test is shown on the right. Each primer set produces a band around 530bp in size.

P[w+] KO lines (red eyes, female infertile) were mated with Tofu KO (white eyes, female fertile). From this cross, (F1) virgin female P[w+] KO / Tofu KO flies (red eyes) were mated with ABKO/ABKO males (white eyes). If recombination between the alleles did not occur, the resulting (F2) female progeny would be P[w+] KO/ABKO (red eyes, female infertile), and Tofu
KO/ABKO (white eyes, female fertile). A recombination frequency less than 50%, would link the locus of the white trans-gene from the mapping lines to Tofu.

![Diagram of recombinant mapping strategy. Flies were segregated based on eye color and individual females were scored for fertility.](image)

The phenotype of Tofu is restored female fertility in the absence of functional BEAF. Tofu is a dominant mutation, so it can be scored in the (F2) progeny where it will be absent or heterozygous. To observe the phenotype of enhanced fertility, we crossed (F1) flies with ABKO, which lack a functional copy of BEAF, and then scored the ability of female (F2) progeny to lay eggs. We began by selecting red and white-eyed female (F2) progeny and isolating them in a single vile with a y-w- male. After one week, the (F2) vials were scored for fertility on the presence or absence of larvae. The intensity of labor and resources required for scoring individual flies this way limited the numbers we could feasibly score, which we held to 200 per cross (100 with white eyes, 100 with red).
If *Tofu* is genetically linked to a *P[w+]* element, then the rate at which the two alleles combine would be less than a random chance of 50%. The crosses would result in (a) red-eyed flies being nearly infertile and (b) white-eyed flies with normal fertility in the scoring generation. The percent of red-eyed fertile should equal the percent of white-eyed infertile. The results, however, were not confirmed by a reciprocal white-eyed infertile/red-eyed fertile ratio but instead white were quite similar to red, indicating a problem with the assay. This may be due to genetic interaction in our flies that we are unaware of, which could decrease the general health and fertility of the flies. White-eyed infertility rates were higher than expected, perhaps indicating *w-* flies are less healthy than *w*+. Evaluating female fertility is tricky because it can be affected by environmental conditions as well as genetics. Fertility rates are easily underestimated. Single female crosses are especially susceptible to environmental conditions because the low number of larvae in the media makes it more susceptible to problems like bacterial or fungal growth. Sometimes the adults die due to the conditions in the vial before they get a chance to lay eggs, thus increasing the observed infertility rate. If this is the case, then the red, fertile recombination rate should be more reliable. It shows recombination has occurred, and should be a minimum estimate of recombination rate if fertility is underestimated. If we look at only the red-eyed flies (Figure 2.6), the data from these crosses implicates cytological regions 27E6, 30C1, 35B1, 35D2, 36E3, and 47A11 as possibly linked regions. Also, it is not possible that all of the indicated regions are linked to *Tofu*. We reasoned that the data from this experiment was unreliable.
We made an early effort to map the left arm of chromosome 2 (2L), because \textit{Tofu} seemed to separate easily from \textit{BEAF} via recombination. Since the \textit{BEAF} gene is in the middle of 2R we reasoned \textit{Tofu} was likely be on the other arm. However our linkage assay suggested that 'far apart' can be closer together than we assumed. A recessive marker mutation speck (\textit{sp}) which is found in the 60B12-60C1 region, approximately half of the chromosome arm distant from \textit{BEAF} at 51C2. This mutation is present in our \textit{ABKO} and \textit{Tofu KO} lines, but not in the NP lines. Thus, \textit{sp} was a convenient marker mutation to get an idea of recombination frequencies within the same chromosome arm. In a recombination cross of \textit{Tofu NP} to a line containing \textit{sp} (Figure 2.7 B) we found the alleles combined at a rate of $49\%$ ($n=305$). In a subsequent cross (Figure 2.7 C) we were able to separate the alleles at rate of $56\%$ ($n=111$). These observation suggested that regions on 2R could contain \textit{Tofu}. 

![Recombination Frequencies](image)

Figure 2.6. Results of recombinant mapping of \textit{Tofu} and the $P[w+]$ lines.
Our results did not support any linked regions, so, we attempted to determine the recombination rate of the two \(P[w^+]\) KO lines nearest to \(BEAF\) (51C2), 49E1 and 53A4. In this cross (Figure 2.7 A) we measured the rate at which the \(P[w^+]\) marker would separate from \(ABKO\). We first mated the \(P[w^+]\) KO lines to \(y\)-\(w^-\) flies. The F1 females from this cross were mated to a balancer line. F2 progeny were genotyped for the WT and \(ABKO\) alleles of \(BEAF\) using PCR. Both regions demonstrated linkage to \(BEAF\) with the closer region, 49E1, separating from \(BEAF\) at a rate of 7.6% (n=79) and 53A4 separating at a rate of 11.3% (n=71). The results confirmed our ability to identify linked regions, confirming the failure of our fertility assay resulted from problems accurately scoring fertility. Unfortunately, we cannot use a PCR based assay to map \(Tofu\) as we would need to know it’s location to design primers.

![Diagram of cross strategies for the separation of alleles from BEAF. A: Recombination of NP6377 and sp. Flies were scored on eye color and sp phenotype. Note: NP6377 contains a \(P[w^+]\) transgene at the BEAF locus. B: Separation of sp allele from the NP6377 allele. Flies were scored on presence of sp phenotype. C: Separation of \(P[w^+]\) transgene. Recombination was scored by PCR genotyping of the ABKO allele.](image)
A Strategy to Confirm the Identity of Tofu Candidates in Cultured Cells

We initially identified 587 protein coding mutations in the genome, which is far too many to test in transgenic flies. As an alternative, we planned to test the candidate genes in a cell culture model. This approach required primary cell cultures from Tofu KO, ABKO and y-w- would also be included as a control. If Tofu is a gain of function mutation due to a protein coding change, the effect may be reversed by silencing the candidate gene with RNAi. We identified “response genes” whose expression is significantly changed when BEAF is knocked down using S2 cell RNA-seq data from Lhoumoud et al 2014. First, we planned to test if the response genes are similarly expressed in y-w- and Tofu KO but differentially expressed in ABKO via qPCR. We then would introduce RNAi of each candidate gene to the Tofu KO cells. Finally, we would monitor response genes via qPCR. A positive result is indicated by response gene levels in Tofu KO that resemble ABKO rather than y-w-. In this experiment, I would attempt to “un-rescue” the phenotype. If a candidate is found, the mutated gene will be cloned into ABKO flies to attempt to rescue loss of fertility.

Generation of Cell Lines.

The fertile phenotype is linked to ovary development, so we attempted to generate a primary cell culture of ovarian stem cells from three lines: y-w- ;Tofu KO; and ABKO. Drosophila ovarian cell culture lines have been established using flies that lack the bag-of-marbles (bam) gene (Niki et al., 2006). These flies have ovaries with arrested development at very early stages, resulting in an abundance of stem cells that form the founding population of the cell culture. Using bam flies was unfeasible for us for two reasons. First we would have to introduce bam mutations into our y-w- ;Tofu KO; and ABKO lines. Second, because BEAF and Tofu phenotypes are implicated in
ovary development, *bam* mutations could obscure the effects of *BEAF* and *Tofu* mutations. Instead, the media was supplemented with the growth factor *dpp* in the form of fly extract. *Dpp* has been shown to block the expression of *bam* and is critical to the maintenance of ovarian stem cells (Chen and McKearin, 2003). Therefore, supplemental *dpp* in the cell media was intended to inhibit the differentiation of ovarian stem cells, allowing them to grow to sufficient numbers. Cells obtained from the cultures were small and tended to form clusters resembling the previously established ovarian somatic stem (OSS) cell cultures (Fig 2.8).

![Figure 2.8. Primary cell culture from the ABKO line at 200x magnification.](image)

Though many wells succumbed to contamination, a few did grow successfully. The *ABKO* line was the most successful presumably because loss of *BEAF* leads to increased numbers of ovarian stem cells, similar to *bam*. Even the ABKO line, however, had an extremely low growth rate, and eventually the line could not be maintained. Interestingly, we found intact tissues could be preserved for weeks in the media. This suggests it is possible to collect tissues for future experimentation.
Reanalysis of SNPs

_Tofu NP_ and _Tofu KO_ were generated by meiotic recombination between _NP6377_ and _ABKO_ chromosomes. An NP6377/ABKO was maintained for many generations before the existence of Tofu was realized, and the homozygous lines _Tofu NP_ and _Tofu KO_ were established. _Tofu_ was transferred between these chromosomes, so nearby variants should be linked to _Tofu_ while the rest of the chromosomes could be shuffled and not show the correct pattern among sequenced lines. According to this line of reasoning, the _Tofu_ mutation should lie in the middle of a variant dense region, so we plotted the _Tofu_ candidate variants on IGV (Figure 2.9). While there are several pockets of variant dense regions, one region on 2R is enriched with variants. It contains (a) 954 of the 1203 candidate SNPs on the second chromosome and (b) 256 of the 587 SNPs that changed protein coding sequence. I refer to this region as the “SNP island”.

![Figure 2.9. SNP density visualization. The SNPs are filtered from top down, the top row represents an unfiltered VCF and the bottom row are candidate SNPs for Tofu. SNP island highlighted in the red box.](image)

The SNP island corresponds to cytological regions 55F8-57C3. The original p-element fly lines we used were outside of this region, so we obtained _P[w+]_ insertion lines at regions within the SNP island 56C6 and 57B16. The results (Table 2.6) were confounding because they suggest
recombination may be encouraged between these regions, rather than suppressed. It is possible the Tofu mutation is being propagated by mismatch repair pathways, resulting in gene conversion (Borts et al., 2000). The process of mismatch repair uses homologous recombination, and mismatch repair proteins can prevent recombination from occurring if the DNA is not sufficiently complementary (Do and LaRocque, 2015). Active transcription can promote recombination (Grimm et al., 1991). Due to the influence of transcription and homology, recombination loci are not random along the genome. Instead, they initiate at specific sites called “hotspots” and are inhibited at “coldspots” (Boulton et al., 1997). An effort was made to assess recombination rates along the Drosophila genome, and the data is illustrated in Figure 2.10 (Comeron et al., 2012). If the Tofu mutation strongly activates transcription, recombination of this region should also increase.

![Figure 2.10](image.png)

Figure 2.10. IGV screenshot of chromosome 2R showing recombination rates (blue), SNP distribution of Tofu candidates (red), and putative PREs (black). Recombination rates were determined by Comeron et al. 2012. The approximate location of BEAF is indicated.

The original focus of the SNP analysis was to identify gain-of-function protein coding mutations that rescue the loss of functional BEAF. Protein coding mutations are predictable by changes in the codons that determine amino acid sequence. Protein coding mutations usually result in
reduced function of the protein, but can result in gain of function as is the case with many oncogenes. A regulatory mutation, causing ectopic expression, is more likely to cause a gain of function. Previously we had not considered this class of mutations. Regulatory mutations are often difficult to detect because they do not alter the gene coding sequence. The limited knowledge of transcription factor binding specificity makes it difficult to predict if an SNP will alter a factor’s binding affinity at that site. Additionally, enhancer-promoter specificity is not well characterized for most genes, making it difficult to predict which genes are affected by a given enhancer. Regulatory regions can act from great distances to alter the expression of target genes, but because this is not well characterized for most genes, only mutations in nearby (cis) regulatory modules (CRM) are feasible for analysis at this time. An intersection of Tofu candidates with a database of CRMs, REDfly, revealed 166 CRMs in the SNP island have Tofu candidate SNPs (Rivera et al., 2019). A mutation that disrupts the function of a silencer would be the most likely to cause a gain of function in which an intact protein gains the ability to function in an atypical context. Therefore, we searched for silencer elements that contain an SNP as regulatory candidates. The best understood silencing elements are polycomb response elements (PREs). A previous study predicted 537 PREs in the genome based on chromatin analysis (Nègre et al., 2011). We found only two predicted PREs with candidate SNPs on the second chromosome, and both are within the SNP island region of 2R (Figure 2.10). One is upstream of the gene Act57B, while the other is upstream of the ribbon gene.

Our lab previously performed a co-immunoprecipitation of BEAF from Drosophila embryos and by mass spectrometry, identified 93 co-immunoprecipitated factors with significant peptide signals. We searched for SNPs in the introns and exons of these proteins and found that 10 of the
potential BEAF interaction partners had an SNP: bl, CG30122, CG4038, hrb27c, lola, me31b, prp19, rib, sdha, and spindly. If these proteins interact with BEAF, they might co-regulate some of the same genes as BEAF, increasing the chance they compensate for loss of BEAF and are stronger candidates for Tofu.

These analyses suggest the mutation of the PRE upstream of the ribbon gene is the strongest Tofu candidate (Figure 2.11). It lies within the SNP dense region of the Tofu candidate SNPs. Furthermore, it was found with BEAF in a co-immunoprecipitation and mass spectrometry analysis. The predicted PRE directly upstream of its promoter is mutated, which could cause more ectopic expression of the protein. The H3K27me3 histone mark is associated with Polycomb silencing and can be found at the PRE upstream and extending through the gene body of ribbon in ChIP-seq data from ovaries (SRX177656). This indicates ribbon is silenced in the wild type ovary, however it is possible it becomes un-silenced in a minority of ovarian cells. A literature search of ribbon revealed it is involved in gonad development (Silva et al., 2016). It has other developmental roles in Malpighian tubules, salivary glands, hindgut, tracheae and head (Jack and Myette, 1999). ChIP-seq data of ribbon from salivary glands revealed a similar DNA binding consensus motif to that of BEAF: CGATA (Cuvier et al., 1998; Hart et al., 1997; Loganathan et al., 2016; Zhao et al., 1995). These findings from the literature further strengthen the idea that ribbon could compete with, cooperate, or replace the function of BEAF.
Figure 2.11. A: IGV screenshot of the candidate PRE upstream of the ribbon gene. The histone mark H3K27me3 associated with polycomb repression was collected from ovaries SRX177656. B: A comparison of BEAF and Ribbon consensus motifs identified by MEME analysis of ChIP-seq peaks. The ribbon motif is from Loganathan et al. 2016. The BEAF motif called from S2 cells SRX3486677.
DISCUSSION

In retrospect, our genomic sequencing did not sample enough populations containing the Tofu mutation. Genome wide association studies in humans use many genomes from individuals to identify SNPs associated with a trait. These studies include a minimum of 2,000 individuals in each group, often exceeding 100,000 (Spencer et al., 2009; Tam et al., 2019). Our analysis was limited to 5 genomes of pooled populations, and this likely limited our ability to screen SNPs. However, similar pooled sequencing strategies have been employed successfully in Drosophila (Blumenstiel et al., 2009). For a better sequencing strategy we could isolate Tofu KO after one recombinant generation of ABKO/NP6377 (using CyO to get isolated chromosomes), and sequence several independently recombined lines to compare to ABKO to find SNPs common to independent Tofu KO lines.

Our original focus for Tofu candidates were SNPs with a predicted protein coding change. The strategy we had in mind was to select several candidates that had moderate or severe protein coding changes and clone the mutant alleles into ABKO flies. Rescue of fertility by a trans-gene would be a conclusive test for Tofu candidates. With 587 potential candidates on chromosome 2 with protein coding mutations, none of them were striking candidates with similar functions to BEAF. For this reason, we employed meiotic mapping to narrow down the list of candidates. Eventually we realized SNP density would indicate the Tofu mutation, which cut the number of protein coding SNPs to 256, still a high number. Looking at potential interacting partners of BEAF, and the inspection of regulatory elements led us to ribbon. If the PRE mutation upstream of ribbon is indeed preventing silencing of the gene, this should be detected by a relative
increase in *ribbon* mRNA levels in *Tofu* flies. No information is known about the expression pattern of BEAF and ribbon in the ovaries, but this data exists for embryos (Kumar et al., 2011). We plan to screen embryos for atypical expression patterns of *ribbon* via FISH in the *Tofu KO* line as compared to the wild type *y-w* using *BEAF* expression as a control. We also plan to screen ovaries where the phenotype manifests. However, there is no data for the expression pattern of *BEAF* or *ribbon* in ovaries, so detection of ectopic expression may be more difficult. If there is no ectopic expression of *ribbon*, we should inspect CRMs in the SNP island for candidate variants. We can use virtual 4C of Hi-C data to look for interactions of candidate CRMs with gene promoters. This may help us select candidates that regulate genes that have characteristics similar to *BEAF*, as *ribbon* does. If the *Tofu KO* line displays ectopic expression of *ribbon*, we will confirm the result by driving ectopic expression of *ribbon* in the ovaries of *ABKO* flies using a *ribbon* trans-gene with a synthetic promoter to look for rescue of fertility (DeLuca and Spradling, 2018).
NOTES


CHAPTER 3. GENOMIC INVESTIGATIONS OF BEAF FUNCTION

INTRODUCTION

Previously, our lab intensively characterized the insulator binding protein BEAF. Here, we present a bioinformatic analysis of relevant publicly available datasets that seek to elucidate BEAF’s role in gene regulation.

BEAF was discovered through the analysis of special chromatin structures (scs & scs’) acting as a boundary at a heat shock puff. The structures were found to contain elements that block the interaction of enhancers with promoters as well as prevent the spread of chromatin domains (Kellum and Schedl, 1991, 1992). This new class of DNA element was named “insulator” after their ability to partition regulatory elements. Insulator function relies on the binding of effector proteins aptly named insulator binding proteins (IPBs). Purification of proteins that binds scs’ lead to the discovery of BEAF and it became one of the first IPBs studied. Microscopy studies have shown that BEAF is present at band/interband boundaries in polytene chromosomes (Hart et al., 1997; Zhao et al., 1995).

Chromatin immunoprecipitation (ChIP) studies for BEAF have been performed several times by a number of labs. Our lab identified 1818 genomic sites based on conservative reproducibility among several embryo ChIP-chip samples using three different antibodies for BEAF (Jiang et al., 2009). ChIP-seq analysis by other labs have reported larger numbers of BEAF peaks ranging from approximately 3000 to 6000 sites (Bushey et al., 2009; Contrino et al., 2012; Lhoumaud et al., 2014; Li et al., 2015; The modENCODE Consortium et al., 2010; Wood et al., 2011). The differing number of sites result in part from data collection factors and peak calling algorithms, but they may also have biological significance. Some of the variability may
result from “indirect peaks” which are thought to result from long distance interaction between BEAF and other factors (Liang et al., 2014). These datasets were collected from various cell types, tissues, and treatments so it is possible that some of the variability is due to gene regulation differences among samples.

Analysis of our lab's conservative loci shows that most BEAF sites are close to gene promoters, a finding that is consistent with other BEAF ChIP datasets (Jiang et al., 2009). Moreover, 85% of BEAF-associated genes are constitutively active (housekeeping) genes (Jiang et al., 2009; Lam et al., 2012). Recent studies in which the scs' element was dissected showed that insulation and activation activities are separable (Maharjan et al., 2020) and promoter-proximal BEAF serves to activate genes (Dong et al., 2020).

Loss of functional BEAF has an effect on development. It causes defects in ovary morphology that lead to near infertility of females and a slight defect in eye morphology (Roy et al., 2007a). These phenotypes are greatly enhanced by the introduction of a dominant negative allele of BEAF (BID) which lacks the DNA binding domain. This BID transgene inhibits the DNA binding of BEAF complexes (Gilbert et al., 2006). Using the BID allele to enhance eye defects, our lab was able to show genetic interactions with developmental transcription factors (TFs), particularly with homeobox proteins in the Antennapedia complex. Heterozygous mutant alleles of these developmental TFs further enhanced the rough eye phenotype in the BID background (Roy et al., 2007b). Other labs have found genetic interactions with developmental pathways such as BMP (Deignan et al., 2016), and Hippo (Jukam et al., 2016).

BEAF itself is expressed ubiquitously (Contrino et al., 2012). Knockdown experiments have shown that BEAF typically serves to keep associated genes active, though a repressive effect was
observed for some genes (Jiang et al., 2009; Lhoumaud et al., 2014). Out of our list of 1,818 BEAF peaks, 1,630 intersect with a 300bp window around the TSS of genes in the dm6 genome. Of these, 650 peaks are located between divergently transcribed genes, making the total number of BEAF-associated genes 2,280. DNA footprinting of BEAF-bound sites suggests CGATA is a consensus binding motif (Hart et al., 1997). One CGATA motif, however, is not enough to ensure BEAF binding, and 139 sites which BEAF reproducibly binds contain no CGATA motif (Jiang et al., 2009). One model to explain lack of consensus motifs is that binding is achieved with the aid of other proteins, either through heterodimerization or facilitated binding at certain sites. An alternative model is these are indirect peaks, resulting from chromatin looping via protein interactions (Liang et al., 2014). Previous findings from yeast-2-hybrid (Y2H) and mass spectrometry of co-precipitating proteins from embryos have identified TFs that may have direct or indirect interactions with BEAF. Among these factors are (a) homeobox containing proteins such as Scr, Bcd, and Abd-B and (b) zinc finger C2H2 proteins: M1BP, row, ribbon, and lola (Maharjan unpublished). Five more homeobox proteins: Ubx, abd-A, Gsc, Irbp18, and bap, have also been reported to physically interact with BEAF using high-throughput assays (Bischof et al., 2018; Shokri et al., 2019). Each of these proteins are potential partners for cooperatively recruiting BEAF to DNA.

The following analyses integrates publicly available and lab-generated datasets from multiple experimental sources to investigate BEAF dynamics between cell types and the response of cells to depletion of BEAF by RNAi mediated KD. In the first half of the study, we investigate the dynamics of BEAF binding in different cell types. We focus on embryonic derived cultured cell lines, S2 and Kc, as a model for differential BEAF binding. The second half
of the study focuses on the effects of BEAF depletion using RNA-seq, PRO-seq, MNase-seq, and Hi-C experimental datasets. We also look for factors and sequence motifs that correlate with BEAF binding sites. From these studies, we find BEAF has a role in gene activation and evidence that it blocks promoter access by the transcription factor, GAF.
METHODS

ChIP-seq

ChIP-seq data was obtained from the ChIP-Atlas database, which provides aligned ChIP-seq data in the form of bigwig files, along with peak files called using MACS2 (Oki et al., 2018; Zhang et al., 2008). The Integrated Genomics Viewer, IGV, was used to show example regions of the data (Robinson et al., 2011). PCA analysis, heatmaps, average plots, and ratio files were made using the DeepTools suite (Ramírez et al., 2016). BedTools was used to compare peak regions (Quinlan, 2014).

Colocalization analysis was done using the ChIP-Atlas online interface. Precise methods for colocalization can be found in Oki et al. 2018. Briefly, MACS2 peaks of individual datasets are scored as High =3, Medium=2, or Low =1 based on binding score. Each score category is independently compared to SRX datasets using a similarity function. The similar categories between sets are multiplied to give a correlation score of 1-9. For example, if the medium peaks (2) of SRX_1 and high peaks of SRX_2 are similar, the correlation score is 6. In the summary table, we show the average score of cell line BEAF datasets and colocalization partners with a score higher than 1, since this is an average between 9 BEAF datasets, the scores range from 0-9. Enrichment analysis takes two genomic region BED files from the user and overlaps ChIP-atlas called peaks with BedTools intersect. Enrichment of overlap between user BED files is determined for each factor using Fisher’s exact test.
Motif Analysis

The FIMO program from the MEME suite was used to find motifs along the genome, and BedTools and Kent utilities were used to create motif bigwig files (Grant et al., 2011; Kent et al., 2010). The AME program from the MEME suite was used to detect enrichment.


Rna-seq

RNA-seq data was obtained from an NCBI project, “Remapping the SRA” (GSE117217), which provides aligned reads in bigwig format, as well as gene count tables used for differential expression analysis. The data was parsed using linux shell commands, and DESEQ2 was used for differential expression analysis (Love et al., 2014). MA plots and PCA analysis of DESEQ2 read counts were done using R functions.

S2 Cell Tissue Culture and RNAi Treatment

Drosophila S2 cells were grown at 25°C in M3 + BPYE medium with 10% fetal bovine serum (FBS) and antibiotic/antimycotic (anti/anti; 100 U/ml penicillin, 0.1 mg/ml streptomycin, 250 ng/ml amphotericin B) from 5 x 105 to 107 cells/ml. Cells at 5 x 106 cells/ml were diluted 5-fold with serum-free M3 + BPYE + anti/anti, and 150 µg dsRNA was added to 15 ml cells in T150 flasks. After incubating 45 min at 25°C, 15 ml of the same medium supplemented with 20% FBS was added to the cells. After 2.5 days, another 150 µg dsRNA was added followed by
30 ml M3 + BPYE, anti/anti and 10% FBS, and the cells were divided into two new flasks. After another 2.5 days, the cells were harvested for isolation of nuclei. Nuclei were isolated as previously described (Kwak et al., 2013), except MD storage buffer (10 mM Tris-HCl pH 8.0, 250 mM sucrose, 1 mM CaCl2, 60 mM KCl, 15 mM NaCl, 1 mM DTT) was used for nuclei for MNase-seq.

Synthesis of dsRNA used a dsDNA template with a T7 RNA polymerase promoter on both ends. The DNA templates were generated by PCR using the following primers: BEAF forward (CTAATACGACTCACTATAGGGAGCAAGGCCAAGACGCTGAG); BEAF reverse (CTAATACGACTCACTATAGGGAGCGCTGATTTGCCCATTTAC); control LacZ forward (GAATTAATACGACTCACTATAGGGAGAGATATCCTGCTGATGAAGC); LacZ reverse (GAATTAATACGACTCACTATAGGGAGACGGAGGAGCTCGTTATCGC).

PRO-seq

Nuclei were isolated and PRO-seq libraries were prepared as previously described (Kwak et al., 2013; Mahat et al., 2016). PRO-seq libraries were sequenced in 50-nucleotide (nt) runs on the Illumina NextSeq500 using standard protocols at the Cornell Biotechnology Resource Center (http://www.BRC.cornell.edu). Raw sequencing reads were processed using the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Illumina adapters were removed with the fastx_clipper tool, and reads were trimmed to 26-mers using fastx_trimmer. Sequencing reads <15 nt were discarded. fastx_reverse_complement was then used to generate the reverse complement of the sequencing reads, which correspond to the sense strand of nascent RNA. Reads were aligned uniquely to the *D. melanogaster* dm5.22 reference genome using Bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2009) with up to two mismatches. Replicates
were correlated and pooled for further analyses. DESEQ2 was used for differential gene expression analysis. DeepTools was used for heatmap correlations and profile plots.

**MNase-seq**

Cells were treated with 1% formaldehyde for 2 min at room temperature, and cross-linking was quenched by adding glycine to 125 mM and placing the cells on ice. Nuclei were isolated as for PRO-seq (Kwak et al., 2013), except they were stored in MD buffer (10 mM Tris-HCl pH 8.0, 250 mM sucrose, 1 mM CaCl2, 60 mM KCl, 15 mM NaCl, 1 mM DTT). MNase digestions were done in 140 µl MD buffer on nuclei from 1.5 x 107 cells with 4000 U MNase (NEB gel units) at room temperature for 30 min, resulting in around 90% mononucleosomes. Reactions were stopped (360 µl 142 mM NaHCO3, 1.42% SDS, 350 mM NaCl, 17.5 mM EDTA, 2.8 mM EGTA), DNA was isolated, and MNase-seq libraries were prepared as previously described (Wei et al., 2012). After 4 PCR cycles, DNA in the size range of 80 to 220 bp (not including adapter sequences) was isolated and sequenced. Paired end read sequencing was done on the Illumina NextSeq500 using standard protocols at the Cornell Biotechnology Resource Center. Illumina adapters were removed with the fastx_clipper tool. Reads were aligned uniquely to the *D. melanogaster* dm5.22 reference genome using Bowtie2 with the parameters “-p 8 -X 250 --no-discardant --no-mixed –no-unal”. After alignment fragments of 120-180bp were selected and trimmed to the central 50bp, these trimmed fragments were used to produce bigwig coverage files. Danpos2 (Chen et al., 2012) was used to assess nucleosome perturbations, and DeepTools was used for plotting.
Enhancer Calling

MACS2 was used to call narrow peaks from the PRO-seq LacZ combined dataset for plus and minus strands. These peak lists were combined for a total of ~8000 peak regions. Peak regions were screened for proximity to a DM3 TSS using BedTools.

HiC

Raw data was accessed from NCBI using the sra_toolkit. Raw reads were processed using the HOMER and bowtie2 for the virtual 4C analysis performed by HOMER (Heinz et al., 2010). Alignment files (sam) were reprocessed using HiCExplorer (Ramírez et al., 2018). HiCExplorer was used to make the summary graphs, aggregation plots, and create log2 ratio matrices. Pygenometracks was used to visualize HiC data (Lopez-Delisle et al., 2020).
RESULTS

Comparison of BEAF ChIP-seq Datasets

We obtained all publicly available Chip-seq datasets for BEAF from the ChIP-Atlas database, as bigwig signal files. This database contains all available ChIP-seq data from NCBI which are aligned and FPKM normalized by the same pipeline producing convenient and reproducible analyses from these datasets. Most BEAF ChIP-seq data is similar, with major binding peaks co-localizing (Figure 3.1). Exceptions include the datasets obtained from larval imaginal wing disks. This data may simply be of low quality due to library preparation, but it is possible it reflects a biological function of BEAF in this tissue. This data is so different from other datasets, and resembled the input controls. Therefore we assume the experiment is flawed, and those data were excluded from further analyses.

Figure 3.1. IGV screenshot of available BEAF ChIP-seq data, grouped by experiment. Background signal for two experiments are label as input.
To examine the differential occupancy of BEAF among datasets and determine any biological significance in the variation, we first performed a principal component analysis (PCA) of ChIP-seq data. Our previous study of BEAF binding showed 85% of BEAF sites are found within a 300bp window around transcription start sites. To inspect the datasets, we looked at the average signal of each dataset within a 300bp window around the TSS of all annotated genes (n = 17650) in the dm6 assembly. This data was used to direct PCA to assess the degree of variance among the datasets. The top 1000 most variable TSS were used to construct the PCA. The analysis showed embryonic derived tissues cluster together, away from the embryonic derived cell lines, Kc and S2. The Kc cell datasets do not cluster well but are clearly separable from the S2 dataset.
The variability in coverage and fragment size of ChIP-seq datasets makes them difficult to compare with one another. If BEAF binding is variable among cells, these differences may be obscured by artifacts of data collection. To circumvent this problem, we assessed BEAF binding intensity at the TSS within a single dataset and then compared TSS ranking between datasets. The results of these comparisons (Figure 3.3) show relative BEAF ChIP-seq signal more variable between individual experiments in different cell types (S2 and Kc) than between the same cell type (Kc_1 and Kc_2), however these differences were not found to be significant (Pearson’s chi-squared test p=0.66). This type of analysis would be greatly improved by experimental replicates, however only single replicates were used in these studies.

Figure 3.2. Principle component analyses of available BEAF ChIP-seq datasets, showing the variability of BEAF binding at the TSS. S2_nt (SRX386677), Kc_nt_1 (SRX085407), Kc_3hr_HE (SRX085408), Kc_48hr_HE (SRX085409.bw), Kc_nt_2 (SRX749021.bw), Kc_HS (SRX749020.bw), Kc_interphase (SRX100252.bw), Kc_mitosis (SRX100253.bw), 0-8hr_embryo (SRX119302.bw), 12-14hr_embryo_r1 (SRX467097.bw), 12-14hr_embryo_r2 (SRX467098.bw), nt: no treatment; HE: 500nM 20-hydroxyecdysone; HS: heat shock 36.5°C for 20 min. Interphase and mitosis: hydroxyurea 15 ng/ml 16hr, nocodazole 2ng/ml 8hr.
S2 and Kc167 are embryonic derived cell culture lines that are commonly used in Drosophila research, so there is a wide pool of data available with which to study these cells. Cell lines should display uniform stable gene regulation relative to tissue and embryo samples which contain many cell types, lowering the precision with which we can analyze their data. For these reasons, we focused on differential BEAF binding between S2 and Kc cell lines.

In light of the insignificant result seen in Figure 3.3, we devised a different method to test the hypothesis that the relative level of BEAF at a gene promoter might be linked to its activation of that gene. First, we compared binding signals between the SRX386677 S2 dataset and the SRX100252 Kc dataset, which are similar in normalized coverage depth, using a log2 ratio. Then plotted the ratio signal in a 300bp window around TSS, to rank TSS by differential signal. Then, we plotted each dataset according to this sorting. The results, shown in Figure 3.4, indicate that there is differential binding of BEAF between the cell types. To select genes with a differential
BEAF signal at the TSS, an arbitrary cutoff of 1.05 log2 ratio was used, resulting in 142 genes in S2 and 148 in Kc cells. TSS regions that displayed differential binding between cell types were more likely to lack a CGATA motif.

Figure 3.4. Heatmaps of BEAF ChIP-seq data sorted by log2ratio of SRX386677 and SRX100252. CGATA motif distribution is shown in the far right panel. Log2ratio plot not shown in figure. Scale bars are omitted as each dataset has different coverage. Purple indicates BEAF binding. S2 enriched TSS are at the top and Kc at the bottom. Dm6 TSS are shown n=17650.

We wanted to investigate the possibility of a correlation between differential gene expression and differential BEAF binding by comparing the transcription profiles of S2 and Kc cells with
respect to these genes. To do this, we leveraged a database of RNA-seq data compiled by NIH (GSE117217). Searching the database, we found six no-treatment datasets for each cell type. We then used DESEQ2 to compare gene counts between the two cell types. To visualize the data, we made an MA-plot (log2 fold change vs mean expression) and highlighted the genes with differential BEAF binding in their respective cell lines (Figure 3.5). In the plot, we can see a correlation between relative BEAF binding and transcription of genes.

Figure 3.5. MA plot (Y axis= Log 2 fold change between samples. X axis = Mean of gene counts among samples) of differential RNA-seq gene counts in S2 versus Kc Cells. Highlighted dots represent genes with differential BEAF ChIP-seq signal at their promoters in S2 (n = 142) (magenta) and Kc (n= 148)(cyan). Dm6 genes: n = 17472. S2 data: (SRX016201, SRX016202, SRX1212543, SRX1212544, SRX1250572, SRX1250573) Kc data: (SRX1032745, SRX1032746, SRX1433880, SRX1433881, SRX1493047, SRX1493048).
Genes that are expressed in most cell types are classified as housekeeping genes because it is thought they are needed for essential cellular processes. A list of 5,534 housekeeping genes was made by analyzing gene expression from Modencode for genes that have low variance between cell types, tissues, and developmental time points (Graveley et al., 2011; Lam et al., 2012). Of the 2,280 BEAF-associated genes, 1,857 are constitutively expressed and considered housekeeping, while 357 are regulated (or developmental) genes. Figure 6 (below) illustrates that housekeeping genes undergo significant (log2 fold change > 2, mean expression > 10) differential regulation, even in similar cell types like S2 and Kc. Despite being constitutively active, housekeeping genes are still subject to regulation, while most have similar expression between the cell types some housekeeping genes have significant differential expression.
Figure 3.6. MA plot of differential RNA gene counts in S2 versus Kc Cells. Highlighted dots represent housekeeping (n = 1,857) (magenta) and developmental (n = 357) (cyan) genes. Dm6 genes: n = 17,472. S2 data: (SRX016201, SRX016202, SRX1212543, SRX1212544, SRX1250572, SRX1250573) Kc data: (SRX1032745, SRX1032746, SRX1433880, SRX1433881, SRX1493047, SRX1493048).

A more extreme comparison of ovary vs brain gene expression (Figure 3.7) shows larger variability of housekeeping gene expression. While most BEAF-associated housekeeping genes are active in both tissues, many demonstrate significant variation. The robust expression of housekeeping genes suggests they have multiple redundant mechanisms that ensure their expression. Dependent upon available factors in a given cell type, many mechanisms likely work in synchrony to adjust the activation level of these genes. BEAF itself is only one of the mechanisms contributing to activation.
BEAF-associated Promoters

Housekeeping genes often have dispersed promoter regions with multiple TSSs, while developmental genes usually have focused promoters and a single TSS (Vo Ngoc et al., 2019). This dispersed promoter architecture would facilitate regulation by multiple factors, leading to robust, redundant expression. A study was done to determine TSS size using multiple datasets that map the location of 5’ mRNAs (Hoskins et al., 2011). To see if BEAF binding was related to promoter size, we extracted 2,660 BEAF-associated promoters from the Hoskins et al. study.

Figure 3.7. MA plot of differential RNA gene counts in female Brains versus Ovary tissue. Highlighted dots represent housekeeping (n = 1,857) (magenta) and developmental (n = 357) (cyan) genes. Dm6 genes: n = 17472. Female brain: (ERX2252496, ERX2252497, ERX2252498) Ovary: (ERX2159569, ERX2159570, ERX2159571).
using S2 cell ChIP-seq data (SRX386677). The analysis revealed BEAF binds a continuum of promoter sizes but is most commonly found at dispersed promoters (Figure 3.8).

Figure 3.8. Line graphs illustrating the relative sizes of promoters across the whole genome (left) and BEAF-associated promoters (right). 39% of all promoters and 25% of BEAF-associated promoters are less than 100bp.

Dispersed promoters generally have a well positioned nucleosome array usually associated with housekeeping genes. DRE and Ohler motifs 1, 6, and 7 are associated with dispersed gene promoters, while TATA, Inr, Ohler motif 10 (MTE), and DPE are often found in focused gene promoters. The TCT motif is found at ribosomal protein coding genes which are considered housekeeping but have a focused promoter (Vo Ngoc et al., 2019). BEAF peaks can be found at many ribosomal protein promoters. We took the list of 2,660 BEAF-associated genes, sorted them by size as determined by Hoskins et al, and split the list into four even sets. We plotted BEAF ChIP-seq data from S2 cells along with the motifs discussed above onto the terminal TSSs of this sorted list, in order to determine (a) where BEAF binds in relation to the TSS and (b) whether or not there is a relationship between BEAF binding and the distribution of these motifs (Figure 3.9).
From this analysis, we see that the location of BEAF binding is related to the size of the promoter region. BEAF-associated promoters range in size from 10 to 608bp, and the distance between the BEAF binding site and the terminal TSS increases with promoter size. BEAF binding is consistently associated with the DRE and Ohler7 motifs.

We looked for enrichment of motifs at BEAF-associated TSS compared to all dm3 TSS using AME. Using Fisher’s exact test with Bonferroni correction we found motifs enriched with adjusted p-values of: DRE 6.72e-226, Ohler 7 motif 3.18e-50, Sry-delta 3.05e-36, and Ohler 6 1.88e-3. It was previously known that BEAF and the transcription factor DREF can bind the DRE motif which contains the CGATA consensus binding motif of BEAF. BEAF and DREF compete for binding DRE motifs, though their binding sites do not always overlap. BEAF does not bind all DREs, only binding if additional CGATA motifs are present. DREF binds all DREs, but not CGATA that is not part of a DRE (Hart et al., 1999).

Figure 3.9. Distribution of BEAF binding and associated motifs at promoters sorted by size, longest promoters at the top. Core promoter motifs that did not show an association are omitted from the figure. n=2660
While we found Ohler 6 to be slightly enriched in BEAF-associated TSS, we did not see a clear pattern in relation to BEAF binding. The Ohler 7 motif seems to be distributed directly upstream of DRE at BEAF-bound promoters and thus may play a role in directing BEAF binding. We also see an association with AT-rich sequences upstream and downstream of BEAF binding sites, which may further define BEAF-associated TSS binding. Sry-delta was shown to physically interact with BEAF to cooperatively activate promoters (Dong et al., 2020). We found this motif is enriched in BEAF-bound loci. While not a core promoter element, enrichment of the Sry-delta motif at BEAF-associated TSS suggests cooperative binding at these sites.

**BEAF Associated Gene Regulation: RNA-seq**

A previous study of BEAF binding revealed 75-90% of associated genes are active in all tissues analyzed. RT-PCR from this analysis suggests BEAF has an activating effect on associated genes (Jiang et al., 2009). There are two RNA-seq datasets of S2 cells after BEAF knockdown. The first was in a study by the Cuvier laboratory (Lhoumaud et al., 2014). The next dataset comes from a high-throughput RNA-seq experiment by NIH in which they attempted to knockdown (KD) transcription factors expressed in S2 cells in parallel (Lee et. al. GSE81221). This experiment attempted the KD of 483 transcription factors including BEAF. The data from both studies are shown in Figure 3.10.
The data collected from these BEAF knockdown experiments indicate misregulation of a small number of genes both positively and negatively. Lhoumaud et al. reported over 2,000 genes were differentially expressed, 57% of which were BEAF-associated. Our inspection of their data revealed low numbers of read counts, which likely contributed to the stochasticity of the data. The data from Lee et al. was part of a high-throughput experiment with no associated publication. Unfortunately, this data displayed only moderate changes from LacZ for most factors analyzed in the experiment. We suspect the high-throughput nature of the experiment allowed for incomplete knockdown from the RNAi treatments. Alternatively, many of the factors may cause only a small change in gene expression, and the sequencing was not of sufficient depth to detect these small changes. Another study of BEAF KD using microarray technology (not shown) reported only six differentially expressed genes which were all downregulated (Schwartz et al., 2012).

Figure 3.10. MA plots of the effects of BEAF RNAi on gene transcription in S2 cells. BEAF-associated genes are highlighted in red. Data from GSE57167 (Lhoumaud et al., 2014) and GSE81221 (Lee et al. 2016). Dm6 genes: n = 17472.
**BEAF Associated Gene Regulation: PRO-seq**

To better understand the effects of BEAF on transcription, we performed an analysis using precision run-on sequencing (PRO-seq). PRO-seq detects actively transcriptionally engaged RNA Pol II with strand-specific nucleotide resolution. The data gives us a snapshot of transcription at the time of isolation. PRO-seq is therefore more precise in measuring active transcription than RNA-seq which can be affected by factors such as mRNAs with long half-lives. The single nucleotide resolution of PRO-seq allows for the analysis of transcription in promoter regions, where RNA pol II can pause before elongation, a key regulatory step in transcription (Mahat et al., 2016).

We performed RNAi mediated knockdown of BEAF in S2 cells, followed by PRO-seq and Mnase-seq, with a knockdown efficiency over 95% (Western blot not shown). Analysis of the PRO-seq data was split between (a) promoter-proximal regions (PR) (maximum read counts in a 50bp window ranging -50 to 150bp from the TSS) to detect changes in pausing and (b) gene body (GB) (200bp downstream of the TSS and 200bp upstream of the gene end) to detect active elongation. We used DESEQ2 to identify differentially expressed genes with a false discovery rate of 0.05. Knockdown of BEAF resulted in 178 differentially expressed genes. 57 PR and 20 GB were upregulated, and 47 PR and 64 GB were downregulated (Figure 3.11 A & A’).
To see how the effects of the knockdown correlate with BEAF-32 binding, we used ChIP-seq data from S2 cells (SRR386677) to rank genes by BEAF binding in a 300bp window around the TSS (Figure B). Using this sorting, we plotted the PRO-seq data. Looking at the data, we can see that genes bound by BEAF tend to have active promoters and knockdown of BEAF tends to weaken this activation (Figure). This activation, though pervasive, was too low to pass the significance thresholds we used for DESEQ2 (FDR=0.05). This is consistent with reports that BEAF aids in gene activation (Dong et al., 2020; Emberly et al., 2008; Jiang et al., 2009; Maharjan et al., 2020). It also suggests BEAF’s role in activation is minor or redundant as less
than 5% of BEAF-associated genes were affected at significant levels. Downregulation of genes after BEAF KD is consistent with a role in promoter activation; however, several genes showed significant upregulation which is consistent with BEAF’s classical function as an insulator binding protein. To understand the mechanisms involved in these upregulated genes and factors that might participate with BEAF in the downregulated genes, we searched for DNA binding factors that might genetically interact with BEAF at these sites. We used ChIP-Atlas enrichment analysis to scan binding regions from 1,714 ChIP-seq datasets classified as “TFs and others” for overlap with differentially expressed genes.

The genes called as significant by DESEQ2 (FDR = 0.05) in the PR and GB gene groups were combined. Upregulated genes (n = 77) were compared to downregulated genes (n = 111) for each treatment (Table S1). A 300bp window around the TSS was used to detect overlap between the two gene groups and peak files from ChIP-Atlas. Downregulated genes were enriched with mof, trx, BEAF, pzg, row, Nup107, M1BP, and ash1. This finding confirms that BEAF binding is correlated with activation by BEAF. Genes upregulated after BEAF KD were enriched in factors involving Polycomb silencing E(z), Pc, Psc, Sce, and Trl (GAF). Mcm2, lilli, ZIPIC, trx and mod(mdg4) were also enriched in the upregulated gene TSSs. The association of Polycomb group related factors with upregulated genes suggests BEAF may have a role as an IBP, blocking the activation of Polycomb regulated genes. Later, we focus on a relationship with GAF as it is known to recruit Polycomb complexes (Mulholland et al., 2003). Interestingly separate replicates of trx, a factor involved in unsilencing Polycomb elements were found to be enriched in both groups (Geisler and Paro, 2015). This further implies a relationship between BEAF and Polycomb mediated gene regulation.
BEAF Affects Enhancer Activity

Enhancers have been shown to recruit RNA Pol II and initiate the process of transcription (Kim et al., 2010; Santa et al., 2010). The ability of PRO-seq to detect nascent RNAs makes it an ideal data source to identify potential enhancers (Wang et al., 2018). We decided to leverage our PRO-seq data to detect potential enhancers that are sensitive to BEAF depletion.

To detect enhancers, we used a peak calling algorithm to find sites with abundant signal in the control replicates; these are sites where RNA polymerase is paused. We then screened for sites at least 1000bp away from an annotated TSS, resulting in 2,039 potential active enhancers. These enhancer candidates were then k-means sorted by the log2 ratio score in a 20bp window around peak centers between BEAF and LacZ RNAi PRO-seq replicates, resulting in three categories: increased (log2 > 1.15), no change, and decreased (log2 < 0.85). As shown in Figure 13, enhancers that decreased in PRO-seq signal after BEAF RNAi had the greatest change, and these sites were more likely to harbor a minor peak from ChIP-seq data.

Figure 3.12. Average plots of potential enhancer sites categorized by response to BEAF RNAi. LacZ and BEAF PRO-seq signal are plotted on these regions along with BEAF ChIP-seq. Potential enhancers were categorized based on response to BEAF KD: Increased n=619, no change n=953 and decreased n=467.
This analysis suggests BEAF is involved in enhancer activation. A study on enhancer-promoter specificity has shown that housekeeping promoters can act as enhancers for other housekeeping promoters (Zabidi and Stark, 2016). It has been established that BEAF interacts with itself (Avva and Hart, 2016; Hart et al., 1997). A recent study from our lab shows that promoter-proximal BEAF can interact at a distance with the TF Sry-delta to activate promoters (Dong et al., 2020). Taken together these findings suggest that BEAF might use distally located enhancers to promote transcription in conjunction with other TFs. Though this analysis excludes promoters, many BEAF mediated enhancer-promoter interaction may also be housekeeping gene promoters acting as enhancers for one another (Zabidi and Stark, 2016).

**MNASE-seq**

Along with PRO-seq, we performed MNase-seq to observe how BEAF depletion affects nucleosome structure. BEAF most often binds housekeeping promoters which have well positioned nucleosome arrays. These arrays are defined by a nucleosome free region (NFR) bracketed by an upstream -1 nucleosome and a +1 nucleosome just downstream of the TSS (Mavrich et al., 2008). The most striking effect of BEAF depletion is on the -1 nucleosome, which increases in occupancy. The +1 nucleosome is not affected much, shifting very slightly into the NFR. The nucleosome array downstream, however, tends to decrease in occupancy. This effect is most pronounced at divergently paired genes (DPGs), which are oppositely oriented genes separated by 100-1000bp using this definition, we find 1955 DPG pairs in the dm3 genome. Strong BEAF peaks are found at 436 DPG pairs. Depending on the distance between DPGs, there may be 1-3 nucleosomes. To better align -1 nucleosomes, the DPGs are split into
four even sets based upon pair distance (Figure 3.13). BEAF may play a role in positioning these central DPG nucleosomes to help coordinate transcription of the paired genes.

**Figure 3.13.** MNase signal on BEAF-bound DPGs. DPGs (n=436) are split into four categories based on distance between genes. LacZ KD combined replicates (blue) and BEAF KD combined replicates (green) are shown. Distance between DPGs in bp: IV: 1000-578, III: 585-448, II: 448-346, I: 346-130.

### Screen for Co-regulator Partners of BEAF

To look for factors that co-regulate BEAF target genes, we returned to the high throughput KD of TFs in the S2 cell dataset (GSE81221). Although many factors included in this experiment displayed only a marginal effect on gene regulation, we searched for factors that demonstrate similar effects on gene regulation as BEAF KD. To do this, we used a list of 102 BEAF activated genes identified by PRO-seq data to profile gene expression across datasets with DESEQ2. We
then used PCA analysis to cluster samples that had an effect similar to BEAF. As shown in Figure 3.14, BEAF KD only had a marginal effect on the target genes, but it was separable from the cluster of control replicates. Factors that deviated from the cluster in the same direction as BEAF may co-regulate the target genes. These factors are summarized in Table 3.1.

Figure 3.14. PCA Plot of RNA-seq after TF knockdown in S2 cells (n=1900). The box indicates factors that exhibit co-regulation with BEAF which are listed in Table 1. BEAF (green) is at the top left corner of the box. The LacZ control replicates (red) are buried in the central cluster. All other datasets (blue). PCA was done using read counts from GSE81221, on 102 genes activated by BEAF according to out PRO-seq analysis.
Table 3.1. List of factors which show a strong effect on BEAF regulated genes. As determined by PCA similarity.

<table>
<thead>
<tr>
<th>Factors similar to BEAF by PCA clustering of RNA-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>az2</td>
</tr>
<tr>
<td>bl</td>
</tr>
<tr>
<td>Cdc5</td>
</tr>
<tr>
<td>eg</td>
</tr>
<tr>
<td>CG10209</td>
</tr>
<tr>
<td>CG10321</td>
</tr>
</tbody>
</table>

This screen likely contains false positives for transcriptional coregulators of BEAF. However, it does provide a list of candidate partners. We can compare this result with other high-throughput screens. By having multiple lines of evidence for genetic and physical interactions, we select candidate factors for physical and genetic interactions with BEAF.

**Screen for Colocalization of BEAF with Transcription Factors**

To look for factors that co-regulate BEAF-associated promoters, we used ChIP-Atlas to inspect the colocalization of BEAF with various TFs. ChIP-Atlas is a data mining suite that contains processed data files of public ChIP-seq datasets including 1,792 datasets of factors that affect transcription in *Drosophila* (*Oki et al., 2018*). We searched for colocalization of factors with BEAF in datasets collected from cell lines. The significant matches are summarized in Table 2. See methods for details about scoring colocalization.
Overall, the results of this and the previous section provide a basis for studies of BEAF involvement in various transcriptional regulatory processes. The presence of other insulator proteins may confer additional stability to the regulation of these genes, accounting for the marginal effects from the loss of BEAF. The presence of many chromatin remodeler subunits, histone modification subunits, and transcription factors suggests that BEAF-associated genes are regulated throughout development and in conditions of cellular stress. The presence of both activating and repressive chromatin remodeling complexes may reflect the dual roles of BEAF as an insulator and activator of transcription.

Table 3.2. Summary of ChIP-Atlas colocalization results. Each ChIP-seq peak file is split into categories based on peak strength (LOW = 1, MID = 2, HIGH = 3) and treated as an independent dataset. Similarity between sets is evaluated and matching set scores are multiplied. A BEAF MID (3) peak set similar to a LOW (1) TF peak set would produce a score of 3. Nine BEAF datasets from cell lines were analyzed for colocalization with other factors and the scores were averaged. Scores range from 0-9. A score of 9 indicates colocalization of strong peaks between datasets. A score less than 1 indicates weak peaks colocalize between some but not all datasets.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Mean Score</th>
<th>Factor</th>
<th>Mean Score</th>
<th>Factor</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ash1</td>
<td>9.00</td>
<td>mof</td>
<td>7.00</td>
<td>Med</td>
<td>0.89</td>
</tr>
<tr>
<td>BEAF-32</td>
<td>9.00</td>
<td>Nup107</td>
<td>7.00</td>
<td>Nup98-96</td>
<td>0.89</td>
</tr>
<tr>
<td>Cap-H2</td>
<td>9.00</td>
<td>Rrp40</td>
<td>7.00</td>
<td>CTCF</td>
<td>0.78</td>
</tr>
<tr>
<td>Chro</td>
<td>9.00</td>
<td>ZIPIC</td>
<td>6.44</td>
<td>fs(1)h</td>
<td>0.44</td>
</tr>
<tr>
<td>Dref</td>
<td>9.00</td>
<td>Cp190</td>
<td>5.33</td>
<td>su(Hw)</td>
<td>0.44</td>
</tr>
<tr>
<td>Pc</td>
<td>9.00</td>
<td>Rrp6</td>
<td>5.33</td>
<td>vtd</td>
<td>0.44</td>
</tr>
<tr>
<td>pzg</td>
<td>9.00</td>
<td>CBP</td>
<td>4.67</td>
<td>CG7099</td>
<td>0.22</td>
</tr>
<tr>
<td>trx</td>
<td>9.00</td>
<td>Iswi</td>
<td>2.00</td>
<td>Lint-1</td>
<td>0.22</td>
</tr>
<tr>
<td>DnaJ-1</td>
<td>7.00</td>
<td>Clamp</td>
<td>1.33</td>
<td>Mad</td>
<td>0.22</td>
</tr>
<tr>
<td>LysE</td>
<td>7.00</td>
<td>msl-2</td>
<td>1.33</td>
<td>mod(mdg4)</td>
<td>0.22</td>
</tr>
<tr>
<td>M1BP</td>
<td>7.00</td>
<td>woc</td>
<td>1.00</td>
<td>shn</td>
<td>0.22</td>
</tr>
<tr>
<td>Mlf</td>
<td>7.00</td>
<td>l(3)mbt</td>
<td>0.89</td>
<td>Smox</td>
<td>0.22</td>
</tr>
</tbody>
</table>

103
The Role of BEAF in Genome Organization

BEAF is often found at TAD borders (Cubeñas-Potts et al., 2017; Rowley et al., 2017). This fact, along with BEAF’s activity related to insulation led to the model that BEAF is an architectural protein that helps shape genome organization. To investigate this activity, Ramírez et al. performed Hi-C experiments after BEAF knockdown in S2 cells. This study found the BEAF binding motif to be enriched at TAD boundaries. It also observes TAD boundaries are often located at constitutively active genes which is consistent with BEAF regulating housekeeping genes. TAD boundaries, however, were not affected by BEAF knockdown (Ramírez et al., 2018). Although the effects of BEAF knockdown were not dramatic, this data can provide insight into BEAF’s role in genome organization.

When we analyzed the data from this experiment, we found that BEAF knockdown results in (a) a larger fraction of short range interactions (less than 20kB) and (b) less long range interactions compared to the GST control knockdown (Figure 3.15). The interactions of specific loci can be gleaned from Hi-C data, giving us virtual 4C data. Virtual 4C on loci containing a BEAF ChIP peak consistently gave smaller windows of local interactions after BEAF KD (not shown). Most genomic interactions are local, but the data suggests BEAF can facilitate local interactions at longer distances.
The results from our PRO-seq analysis suggested BEAF blocks GAF from activating certain TSS. We hypothesized that depletion of BEAF would allow GAF to contact TSS formerly bound by BEAF. This would increase the contact frequency between BEAF-associated TSS and GAF binding sites. To test this hypothesis, we looked for interactions between BEAF (SRX386677) and GAF (SRX183520) ChIP-seq peaks from S2 cells in the GST and BEAF knockdown Hi-C replicates. We found both an increase and decrease in contacts between BEAF and GAF associated loci, but a larger number of contacts showed an increase after BEAF KD. Furthermore, contacts that increased displayed a greater change than those that decreased (Figure 3.16).

Figure 3.15. Summary plots of Hi-C data. The left plot compares the relative fractions of Hi-C contacts BEAF KD results in more short range contacts and fewer long range contacts on average. The right plot validates the quality of the samples from the experiment. This shows all replicates have similar sequencing quality, but BEAF KD results in fewer Hi-C contacts.
TAD architecture seems to be more dependent on the transcriptional state of genes than the binding of particular proteins. Active genes interact preferentially with other active genes, and the inactive genes interact with inactive genes (A & B compartments). BEAF appears to strengthen gene activation, but it is not essential for gene expression of most of its targets. Additionally, BEAF is capable of blocking non-specific enhancer-promoter interactions. BEAF appears to be facilitating specific contacts between genomic regions, while blocking non-specific contacts. This activity is likely due to BEAF-BEAF interactions and interactions with other factors which serve to bring promoters together to increase transcription (Dong et al., 2020). Housekeeping promoters have been shown to act as enhancers for other housekeeping promoters.
in artificial constructs (Zabidi and Stark, 2016). A primary function of BEAF may be to help aggregate these promoters for stable transcription. As we can see in Figure 3.17 these dual activities result in both gain and loss of contacts along the genome when BEAF is depleted. The gain or loss of contacts can be attributed to other transcription factors such as GAF that are free to interact with partners, once they are unencumbered by the constraints of BEAF.
Unfortunately not enough is known about enhancer-promoter interaction pairs to predict all the perturbations caused by loss of BEAF. As we can see in Figure 3.17, there are many perturbations which BEAF-GAF interactions do not account for. To predict these disruptions, we would need (a) a complete interaction map of enhancers and promoters (b) their...
transcriptional state in the relevant cell type and (c) the factors that activate transcription at these sites. Since BEAF has separate activating and insulating activities we would need to take into account those activities at each individual BEAF binding site. However, we can use this data to inspect the interactions of individual BEAF sites and contacts between interaction candidates to detect insulation activity and enhancer-promoter interaction as a basis for further study.
DISCUSSION

The findings presented in this study give us insight into the dynamic nature of gene expression and how insulator binding proteins such as BEAF play a role in this process. By comparing ChIP-seq data on the binding of BEAF in different cell contexts we can see that target loci remain relatively consistent among cell and tissue types (Figures 3.1 & 3.2), but the strength of binding for particular sites varies between cell types and experimental conditions (Figure 3.3). Binding sites that vary the most between the S2 and Kc cell datasets are depleted in the CGATA consensus binding motif (Figure 3.4). This suggests BEAF may be recruited to additional sites by cell specific transcription factors that facilitate BEAF binding to sub-optimal sites. Figure 3.5 shows the correlation between differential BEAF sites in S2 and Kc cells and gene activation in their respective cell types. This indicates the strength of binding of BEAF at the TSS is linked to the level of transcriptional activation for that gene.

Recent studies of BEAF confirm that it has a role in promoter activation, which is separable from it’s ability to insulate genes from enhancers and chromatin effects (Dong et al., 2020; Maharjan et al., 2020). This finding may explain the repression and activation of genes seen in RNA-seq data from BEAF depleted cells (Figure 3.10). RNAi treatment can have false positives due to indirect effects resulting from changed expression of a TF in response to KD of the target protein which in turn alters further gene expression. Association of BEAF ChIP seq with downregulated genes argues against indirect effects on these genes. Analysis of BEAF-GAF interactions argues that upregulated genes identified by PRO-seq are direct effects of BEAF blocking GAF access to certain promoters. Some genes depend on the insulation properties of BEAF to prevent aberrant activation or silencing, while others benefit from it’s activating
properties. This results in most BEAF-associated genes demonstrating stable expression among cell types. However, as can be seen in Figures 3.6 & 3.7 many of these genes are regulated among cell types. BEAF seems to be ubiquitously present at most of it’s target genes, but the activity of these genes is influenced by other cell specific transcription factors.

BEAF-associated promoters typically have a dispersed architecture with multiple TSS (Figure 3.8). The BEAF binding sites at these promoters often contain the DRE and Ohler 7 core promoter elements flanked by A-T rich sequences (Figure 3.9). A-T rich sequences were reported to target the NSL complex to TATA-less promoter, indicating a relationship with BEAF at these sites (Lam et al., 2019). Two factors associated with the NSL complex were found to colocalize with BEAF (Table 3.2). Mof is a common subunit of non-specific lethal (NSL) complex and the related male-specific lethal (MSL) complex (Sheikh et al., 2019). Nup-98-96, is a nuclear pore complex protein that was shown to interact with NSL members and be required for the expression of certain homeobox genes (Pascual-Garcia et al., 2014). MBD-R2 was indicated as a co-regulator of BEAF KD sensitive genes in the RNA-seq genetic screen of S2 cell TF KD (Table 3.1). MBD-R2 is a subunit of NSL and is important for recruiting the complex to DNA (Sheikh et al., 2019). Msl-2 a unique subunit of MSL, was also shown to colocalize with BEAF indicating a relationship with both complexes, perhaps through interaction with the shared unit mof. Both of these complexes are lysine acetyltransferases involved in gene activation. MSL is involved in *Drosophila* X chromosome activation for dosage compensation in males, while NSL has been shown to be involved in housekeeping gene regulation (Lam et al., 2012; Sheikh et al., 2019). Connection to the MSL complex is interesting considering the disruptive effect on X chromosome structure seen in BEAF knockout flies (Roy et al., 2007a).
Homeobox proteins also bind AT rich sequences and often heterodimerize which further confounds attempts to predict binding regions. These factors are responsible for the body plan and tissue development and several are present in any given cell type. Thus many different homeobox complexes likely bind the same sites, but have different influence over transcription (Bürglin and Affolter, 2016). Several homeobox proteins have been reported to physically interact with BEAF. Homeobox proteins are likely to be among the factors which modulate housekeeping promoter activity along with BEAF. This type of collaboration has been proposed between other homeobox protein and M1BP, which strongly colocalizes with BEAF (Zouaz et al., 2017). ChIP-seq has been done on very few of the 103 factors that make up this group, only Abd-B, Ubx, and bcd were included in the ChIP-Atlas database. All of these showed colocalization with BEAF at hundreds of sites, but only bcd appeared enriched (in embryo tissue not shown). An attempt was made to connect homeobox transcription with BEAF binding by inspecting genes downregulated by Ap, the only homeobox factor differentially expressed in S2 vs Kc cells, however the limited effect by the knockdown of Ap precluded any meaningful results. Though no conclusive results with these factors were found, the connection of BEAF with homeobox proteins was an inspiration for many of the studies conducted in this chapter.

Our PRO-seq analysis supports the model that BEAF aids in promoter activation (Figure 3.11), while blocking activation of promoters by other factors such as GAF. The identification of potential enhancers that also respond to BEAF KD (Figure 3.12) indicates a role in enhancer-promoter communication which was supported recent work (Dong et al., 2020). Screens of high-throughput datasets (Figure 3.9, Tables 3.1 & 3.2) can identify candidates for interacting partners
of BEAF, as well as factors whose binding is blocked by BEAF. By investigating these candidates, we will further illuminate the role BEAF has in gene regulation.

The marginal activation effects BEAF demonstrates on most of its target genes suggests it is not a primary activation factor of them. Virtual 4C analysis from the BEAF KD Hi-C experiment and in-vivo DNA looping experiments have shown BEAF can facilitate short range interactions. This, together with BEAF’s ability to act as a chromatin boundary suggests that it helps neighboring genes and enhancers form active sub-compartments for stable transcription. Aggregation plots of BEAF-GAF ChIP-seq peaks show contacts are disorganized after BEAF KD. This result can explain the genes which are upregulated after BEAF KD. These genes were enriched in GAF and Polycomb group proteins. GAF is known to recruit the Polycomb group silencing complexes to PREs (Ogiyama et al., 2018), but also can activate transcription (Chopra et al., 2008). A recent study has shown depletion of BEAF results in spreading of the Polycomb associated repressive histone mark H3K37me3, and resulted in decreased signal at GAF/dCTCF sites. This study also demonstrated BEAF dependent looping to GAF sites using Hi-C data (Heurteau et al., 2020). These findings indicate a relationship between BEAF and GAF in regulating Polycomb mediated silencing.

It is likely that most of these enhancer-promoter associations BEAF facilitates are in fact housekeeping promoter-promoter interactions which function to enhance the transcription of one another. These stable active compartment groups are likely further regulated by cell specific enhancers which are governed by developmental transcription factors such as the homeobox and Polycomb group proteins. In this way, the cell is able to fine tune gene expression among genes with a similar regulatory scheme with BEAF stabilizing the contacts and aiding the transcription.
This would allow for constitutive expression of these genes with a mechanism for regulation by developmental factors.
NOTES


59. Wood, A.M., Van Bortle, K., Ramos, E., Takenaka, N., Rohrbaugh, M., Jones, B.C.,


(MACS). Genome Biol. 9, R137.


63. Zouaz, A., Auradkar, A., Delfini, M.C., Macchi, M., Barthez, M., Ela Akoa, S.,
and AbdA collaborate with the transcription pausing factor M1BP to regulate gene
transcription. EMBO J. 36, 2887–2906.
CHAPTER 4. CONCLUSIONS

Flies homozygous for a null *BEAF* mutation display developmental defects in eye and ovary formation (Gilbert et al., 2006; Roy et al., 2007a). Eye defects, though minor, were enhanced by mutations in homeobox proteins and other developmental TFs (Roy et al., 2007b). The ovary defects lead to female infertility. These phenotypes implicate BEAF as a participant in developmental regulation. Despite this, we find that BEAF mainly binds to and activates housekeeping promoters, though the loss of BEAF does not significantly alter expression of most of its target genes. Indeed, the viability of BEAF-KO flies suggests that there are alternate mechanisms to keep housekeeping genes active in the absence of BEAF.

The existence of the *Tofu* mutation further suggests that BEAF’s role in activating key genes needed for ovary development is likely redundant. It is likely that in the absence of BEAF, most of these genes are able to achieve near normal regulation. In certain cell contexts, BEAF binding sites may be occupied by other TFs that share similar binding motif affinities such as DREF, Optix, and ribbon (Hart et al., 1999; Loganathan et al., 2016; Noyes et al., 2008). Ectopic expression of such a TF might activate many of BEAF’s target genes that are needed for ovary development, especially in the absence of competition by BEAF. Identification of *Tofu* might help us identify these key target genes and their importance in ovary development.

Meta-analysis of ChIP-seq data tells us that BEAF demonstrates variable binding among cell types, but only at a subset of genes. Some of this differential binding is likely due to experimental variation, but gene expression analysis reveal differential binding at certain genes has biological implications. Correlation of differential BEAF binding in different cell types with RNA-seq data shows BEAF binding corresponds to gene activation. PRO-seq analysis revealed
BEAF enhanced the activation of most of its target genes, while blocking activation at a subset of genes. The TSS of the genes where BEAF blocks activation were enriched in subunits of Polycomb group complexes and GAF. This indicates a role in development by coordinating polycomb silencing in conjunction with GAF. MNase-seq data indicated BEAF lowers the occupancy of the -1 nucleosome of the NFR region around the TSS. This may be an important mechanism for promoter activation through RNA Pol II recruitment. Hi-C data shows that loss of BEAF causes changes in genome conformation, with contacts being both gained and lost. Analysis of the average change in total contacts after BEAF depletion indicates BEAF is helping to separate genes by inhibiting short-range interactions (less than 20kB), which are favored due to proximity and facilitates long range interactions. However, closer inspection reveals increased and decreased changes in contacts at both short and long ranges. Presumably, these changes are the result of a lost contacts mediated by BEAF and a gain of contacts mediated by other TFs (e.g. GAF) which were blocked by BEAF.

Together these observations lead to a model in which BEAF is functioning to stabilize proper enhancer-promoter contacts, thereby aiding housekeeping gene transcription. Most of these enhancers are also housekeeping gene promoters that are able to act synergistically to elevate transcriptional activation (Corrales et al., 2017; Zabidi and Stark, 2016). These housekeeping gene clusters may be modulated by developmental TFs which can recruit additional enhancers to the active compartment. This interaction could be facilitated genetically by developmental TFs binding to enhancers and housekeeping promoters. Concurrently, BEAF blocks the interaction of enhancers and promoters that do not belong in the compartment but would otherwise interact due to proximity and transcriptional state (Gilbert et al., 2006; Roy et
Therefore, BEAF might help to organize the enhancer-promoter interactions within housekeeping clusters. To add to this model, BEAF can be phosphorylated and has an unstructured middle domain (Avva and Hart, 2016; Hart et al., 1997). These characteristics fit the model of proteins that participate in multivalent interactions that facilitate phase separation (Dignon et al., 2020). It is therefore possible that phosphorylated BEAF participates in and may facilitate active transcription hubs by liquid-liquid phase separation interactions (Boehning et al., 2018; Hnisz et al., 2017).

The investigations presented in this manuscript have contributed to past publications (Maharjan et al., 2020; Shrestha et al., 2018). Motif analysis contributed to the original manuscript of our recent publication indicating interactions between Sry-delta and BEAF (Dong et al., 2020). Motif analysis will also contribute to ongoing work regarding BEAF promoter activity. The PRO-seq and MNase-seq experiments are part of a forthcoming publication involving BEAF’s interactions with the PBAP chromatin remodeling subunit, polybromo. Portions of the genomics chapter will be included in a forthcoming publication.

**FUTURE DIRECTIONS**

Once the identity of Tofu is confirmed, we can better understand the mechanisms of BEAF’s role in ovary development. We should identify genes coregulated by BEAF and Tofu to dissect how their expression governs ovarian cell differentiation. Analysis of recent *Drosophila* ovary single cell RNA-seq studies could show which cell types are likely to be affected by the misregulation of these genes. Additionally this dataset might reveal information about *BEAF* expression in the ovarian cells which would further elucidate how BEAF affects ovary development (Jevitt et al., 2020; Rust et al., 2020).
In order to gain a better understanding of BEAF binding dynamics, more ChIP-seq studies should be performed. A comparative study of BEAF binding in several cultured cell lines would give us insight into its role in developmental regulation of housekeeping genes. With this data we could clearly identify differentially bound genes. This would enable us to better leverage public datasets to answer questions about BEAF function. We can test if the level of BEAF binding correlates with the level of gene expression. We can also look for evidence of Polycomb regulation between these sites.

We can dissect BEAF’s role in insulation by knocking down BEAF and performing ChIP-seq on GAF, mod(mdg4), and other factors predicted to be blocked by BEAF. A redistribution of these factors to formerly BEAF-bound sites after BEAF KD would provide strong evidence these factors are blocked by BEAF. In a similar strategy, we can knock-in TFs that we suspect may facilitate BEAF binding and evaluate the change in BEAF binding by ChIP-seq. Candidates for knock-in include homeobox proteins and NSL/MSL subunits.

Enhancers predicted by PRO-seq can be tested in genetic constructs to answer questions related to BEAF’s role in enhancer-promoter communication. We can screen for BEAF mediated enhancer-promoter interactions using luciferase reporter “looping” assays already developed in our lab (Dong et al., 2020). Identified enhancers may be characterized for spatio-temporal expression using transgenic reporter assays (Kvon, 2015). Virtual 4C of identified enhancers could also help determine which promoters they interact with. We can then test these for enhancer-promoter interactions using our looping assay.

These proposed studies will provide insight into the role of BEAF in gene regulation and chromatin organization. Identifying Tofu and the key genes regulated by BEAF in the
developing ovary can inform us about the cell types and cellular processes in which BEAF plays a critical role. A ChIP-seq study of BEAF in various cell types would provide dataset that can serve as a model for BEAF mediated regulation in development. Knock-in of TFs can augment this model and evaluate if BEAF binding can be facilitated by the activation genes via other factors. Knockdown of BEAF and ChIP-seq of potentially blocked TFs can reveal mechanisms of gene regulation by BEAF-dependent insulators. Finally, identifying BEAF-dependent enhancer-promoter interactions can further elucidate BEAF’s role in promoter activation and chromatin organization.
NOTES


Table S1. Chip-Atlas enrichment in TSS of genes identified by PRO-seq and DESEQ2 with an FDR less than 0.05. Only S2 datasets with a log 2 q-value greater than 1.5 are shown. Trl encodes GAF.

<table>
<thead>
<tr>
<th>SRA ID</th>
<th>Antigen</th>
<th>Num of peaks</th>
<th>BF_UP</th>
<th>BF_DN</th>
<th>log p-val</th>
<th>log q-val</th>
<th>Fold Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRX146999</td>
<td>mof</td>
<td>11630</td>
<td>32/77</td>
<td>82/111</td>
<td>-4.99663</td>
<td>-3.1295</td>
<td>0.562559</td>
</tr>
<tr>
<td>SRX1794235</td>
<td>trx</td>
<td>5288</td>
<td>35/77</td>
<td>89/111</td>
<td>-5.9965</td>
<td>-4.02463</td>
<td>0.566905</td>
</tr>
<tr>
<td>SRX386677</td>
<td>BEAF-32</td>
<td>4229</td>
<td>32/77</td>
<td>81/111</td>
<td>-4.69738</td>
<td>-2.88824</td>
<td>0.569505</td>
</tr>
<tr>
<td>SRX326967</td>
<td>pze</td>
<td>8641</td>
<td>34/77</td>
<td>86/111</td>
<td>-5.24111</td>
<td>-3.34179</td>
<td>0.569918</td>
</tr>
<tr>
<td>SRX326966</td>
<td>row</td>
<td>10951</td>
<td>30/77</td>
<td>74/111</td>
<td>-3.70524</td>
<td>-2.15438</td>
<td>0.584416</td>
</tr>
<tr>
<td>SRX6685747</td>
<td>Nup107</td>
<td>6967</td>
<td>30/77</td>
<td>72/111</td>
<td>-3.22496</td>
<td>-1.75577</td>
<td>0.600649</td>
</tr>
<tr>
<td>SRX326966</td>
<td>Nup107</td>
<td>6576</td>
<td>30/77</td>
<td>68/111</td>
<td>-2.51604</td>
<td>-1.3661</td>
<td>0.635982</td>
</tr>
<tr>
<td>SRX1794234</td>
<td>trx</td>
<td>6515</td>
<td>41/77</td>
<td>91/111</td>
<td>-4.40117</td>
<td>-2.71013</td>
<td>0.649493</td>
</tr>
<tr>
<td>SRX084579</td>
<td>ash1</td>
<td>11364</td>
<td>43/77</td>
<td>82/111</td>
<td>-1.91736</td>
<td>-1.03443</td>
<td>0.755939</td>
</tr>
<tr>
<td>SRX128301</td>
<td>ash1</td>
<td>10873</td>
<td>43/77</td>
<td>82/111</td>
<td>-1.91736</td>
<td>-1.03443</td>
<td>0.755939</td>
</tr>
<tr>
<td>SRX3011244</td>
<td>Trl</td>
<td>9711</td>
<td>48/77</td>
<td>44/111</td>
<td>-2.53202</td>
<td>-1.37002</td>
<td>1.57261</td>
</tr>
<tr>
<td>SRX3011243</td>
<td>Trl</td>
<td>9152</td>
<td>48/77</td>
<td>42/111</td>
<td>-2.95623</td>
<td>-1.61507</td>
<td>1.6475</td>
</tr>
<tr>
<td>SRX027830</td>
<td>trx</td>
<td>7535</td>
<td>35/77</td>
<td>29/111</td>
<td>-2.11191</td>
<td>-1.12803</td>
<td>1.73981</td>
</tr>
<tr>
<td>SRX3011241</td>
<td>Trl</td>
<td>7877</td>
<td>37/77</td>
<td>29/111</td>
<td>-2.52163</td>
<td>-1.3661</td>
<td>1.83923</td>
</tr>
<tr>
<td>SRX3011242</td>
<td>Trl</td>
<td>8040</td>
<td>40/77</td>
<td>30/111</td>
<td>-3.15293</td>
<td>-1.72137</td>
<td>1.92208</td>
</tr>
<tr>
<td>SRX3011250</td>
<td>E(z)</td>
<td>6204</td>
<td>32/77</td>
<td>24/111</td>
<td>-2.25516</td>
<td>-1.21038</td>
<td>1.92208</td>
</tr>
<tr>
<td>SRX3011246</td>
<td>Trl</td>
<td>6305</td>
<td>34/77</td>
<td>24/111</td>
<td>-2.88693</td>
<td>-1.57264</td>
<td>2.04221</td>
</tr>
<tr>
<td>SRX097617</td>
<td>lilli</td>
<td>6028</td>
<td>43/77</td>
<td>82/111</td>
<td>-1.91736</td>
<td>-1.03443</td>
<td>0.755939</td>
</tr>
<tr>
<td>SRX183520</td>
<td>Trl</td>
<td>5848</td>
<td>22/77</td>
<td>44/111</td>
<td>-2.08863</td>
<td>-1.11655</td>
<td>2.26531</td>
</tr>
<tr>
<td>SRX475424</td>
<td>Pc</td>
<td>6107</td>
<td>21/77</td>
<td>13/111</td>
<td>-1.94231</td>
<td>-1.03626</td>
<td>2.32867</td>
</tr>
<tr>
<td>SRX050595</td>
<td>Mcm2</td>
<td>4263</td>
<td>23/77</td>
<td>14/111</td>
<td>-2.31484</td>
<td>-1.25452</td>
<td>2.36827</td>
</tr>
<tr>
<td>SRX3404015</td>
<td>Trl</td>
<td>4466</td>
<td>20/77</td>
<td>12/111</td>
<td>-2.01487</td>
<td>-1.07716</td>
<td>2.4026</td>
</tr>
<tr>
<td>SRX027824</td>
<td>Pc</td>
<td>5499</td>
<td>29/77</td>
<td>17/111</td>
<td>-3.24633</td>
<td>-1.75577</td>
<td>2.45913</td>
</tr>
<tr>
<td>SRX447393</td>
<td>ZIPIC</td>
<td>3409</td>
<td>18/77</td>
<td>10/111</td>
<td>-1.94497</td>
<td>-1.03626</td>
<td>2.59481</td>
</tr>
<tr>
<td>SRX3404051</td>
<td>Trl</td>
<td>4202</td>
<td>20/77</td>
<td>11/111</td>
<td>-2.31544</td>
<td>-1.25452</td>
<td>2.62102</td>
</tr>
<tr>
<td>SRX3404027</td>
<td>Trl</td>
<td>3321</td>
<td>17/77</td>
<td>9/111</td>
<td>-2.03541</td>
<td>-1.09403</td>
<td>2.72294</td>
</tr>
<tr>
<td>SRX3404019</td>
<td>Trl</td>
<td>3998</td>
<td>19/77</td>
<td>10/111</td>
<td>-2.37709</td>
<td>-1.26146</td>
<td>2.73896</td>
</tr>
<tr>
<td>SRX191912</td>
<td>mod(mdg4)</td>
<td>6785</td>
<td>31/77</td>
<td>16/111</td>
<td>-4.12299</td>
<td>-2.47146</td>
<td>2.79302</td>
</tr>
<tr>
<td>SRX191911</td>
<td>mod(mdg4)</td>
<td>6427</td>
<td>24/77</td>
<td>12/111</td>
<td>-3.19083</td>
<td>-1.73387</td>
<td>2.88312</td>
</tr>
<tr>
<td>SRX027823</td>
<td>Pc</td>
<td>3727</td>
<td>21/77</td>
<td>6/111</td>
<td>-4.44316</td>
<td>-2.73093</td>
<td>5.04545</td>
</tr>
<tr>
<td>SRX027828</td>
<td>Psc</td>
<td>3794</td>
<td>11/77</td>
<td>3/111</td>
<td>-2.39503</td>
<td>-1.26826</td>
<td>5.28571</td>
</tr>
<tr>
<td>SRX4520673</td>
<td>Hmr</td>
<td>2215</td>
<td>9/77</td>
<td>2/111</td>
<td>-2.08464</td>
<td>-1.11655</td>
<td>6.48701</td>
</tr>
</tbody>
</table>
LIST OF REFERENCES


129


135


Palikyras, S., and Papantonis, A. Modes of phase separation affecting chromatin regulation. Open Biol. 9, 190167.


VITA

J. Keller McKowen III was born in Baton Rouge, Louisiana. His interest in Biology was cultivated by his mother, a veterinarian, and his father who raised cattle and took Keller hunting as a boy. He has always had a passion for creating and decided to pursue a career in science at a young age. Keller attended Louisiana State University as an undergraduate, where he had the opportunity to conduct research in a number of laboratories. As an undergraduate, he worked in Craig Hart’s laboratory for four years. There he gained a background in molecular biology techniques, *Drosophila* husbandry, and insulator function. After graduation, he worked briefly as an aquaculturalist at A&E Testing, Inc. in Baton Rouge. In his graduate studies, Keller rejoined the Hart lab with a focus on genomics.