The Dissertation Committee for Moon Draper certifies that this is the approved version of the following dissertation:

Transcriptional Regulation of the Shaker Homolog Kv3

Committee:

Nigel S. Atkinson, Supervisor

Alan Lloyd

S. Martin Shankland

Harold Zakon

Bing Zhang

Transcriptional Regulation of the Shaker Homolog Kv3

by

Moon Draper, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2005

Dedication

Pour ma amour, qui m'a délivrer

Quand le ciel bas et lourd pese comme un couvercle
Sur l'esprit gemissant en proie aux longs ennuis,
Et quie de l'horizon embrassant tout le cercle
Il nous verse un jour noir plus triste que les nuits;

Quand la terre est changeee en un chacot humide,
Ou l'Esparance, comme une chauve-souris,
S'en va battant les murs de son aile timide,
Et se cognant la tete a des plafonds pourris;

Quand la pluie etalant ses immenses trainees
D'une vaste prison imite les barreaux,
Et qu'un peuple muet d'infames araignees
Vient tendre ses filets au fond de nos cerveaux,

Des cloches tout a coup sauntent avec furie
Et lancent vers le ciel un affreux hurlement,
Ainsi que des esprits errants et sans patrie
Qui se mettent a geindre opiniatrement.

-Et de longs corbillards, sans tamours ni musique,
Defilement lentement dans mon ame; l'Espoir,
Vaincu, pleure, et l'Angoisse atroce, despotique,
Sur mon crane incline plante son drapeaur noir.

-Baudelaire

Acknowledgments

This is the product of the efforts of many people. I would like to thank a few in particular;

My advisor, Dr. Nigel Atkinson, for his guidance and arguments—nothing serves to force reasoning better than facing persistent doubt.

My committee, for their patience.

My lab mates. It's your turn to do the dishes.

And to a long string of assistants who didn't know better than to stay away: Alex Porter Few, Jessica Turner, Susan Young, Heather Zidow and Tim Hofstra.

Alex's inspiration was the very fuel of many of the initial ideas presented here. Jessica's tenacity and integrity led to the first crucial break in the cloning. Susan, who worked with me over the course of three years, was (and remains) one of the most remarkable people that I have known. She patiently stayed the course over many frustrating times, corrected my errors kindly, and tolerated my goofiness with grace.

Heather and Tim were my latest victims. The vast majority of the focus needed to finish came under their watch. I shall always be indebted.

To my love, I will be joining you soon.

Transcriptional Regulation of the Shaker Homolog Kv3

Publication No._____

Moon Draper, PhD

The University of Texas at Austin, 2005

Supervisor: Nigel S. Atkinson

The sequencing of entire genomes has brought into light the code of life, and along with it, many questions. Comparisons between whole genomes revealed that far less material is dedicated to sequencing proteins than previously believed and that the commonality between the genes that different species have is far greater than suspected. Humans share with mice, nearly every gene. The key differences between the two organisms lay in the transcriptional control regions that dictate the expression of those genes.

Kv3 is a member of the Shaker cognate family of potassium ion channels found in Drosophila. These channels are involved in establishing the membrane voltage polarity in excitable tissues. The regulatory region for Kv3 was cloned from genomic DNA and core promoter elements identified. Two distinct promoters were mapped. The same genomic region was cloned from other species of fruit flies and sequenced. These were then used as input for bioinformatics applications to find regions of conserved sequences. Seven distinct blocks of sequence were found. It was reasoned that these are conserved

through some functional constraint against variation. The function of these regions was tested in vivo and discrete regions were shown to regulate expression in the central and peripheral nervous systems. A specific element was found that directs expression in the antennomaxillary sensory structures. A fourth member of the Shaker cognate family was identified and found closely related to Kv3 coding sequences.

Table of Contents

List of Figures

INTRODUCTION

Chapter 1: The junk in the genome

> "I have a hunch that the unknown sequences of DNA will decode into copyright notices and patent protections." —Donald E. Knuth, Stanford lecture series, 2003.

THE AGE OF THE GENOME

We are at the dawn of a new age in science. Just as advances in physics ushered in the atomic age of the 20th century, the Human Genome Project, with its initial release in 2001, might likely make the 21st century the Age of the Genome. Evidence arising from the analysis of the genomic sequences will have an impact on nearly every aspect of biology from molecular studies to population ecology. The genome project has spawned advances in such diverse interests as criminology, anthropology, and sociology. From medical advances to political debate, the implications of knowing the code for life are astounding. However, as with most matters of science, the answers found in the genome have given rise to many questions. What is the function of the non-coding sequences that make up the majority of the DNA in a genome? How is the decision made to express a gene and can this be construed from the genomic sequence? Are we solely the product of the genes housed in our genome, or are there other parts that will reveal themselves as having a greater role in determining who and what we are? Will the discovery of the function of these other sequences change how we describe a gene? In short, the next horizon after the annotation of the genes throughout the genomes is the elucidation of systems and sequences that regulate the expression of these genes.

The central dogma of gene expression that gives rise to proteins is that it occurs in

two steps: the reading of the gene from the chromosome to produce a messenger RNA (mRNA) and the subsequent translation of that mRNA in protein synthesis. In the process of translation, the nucleotide code of mRNA is used to direct the arrangement of amino acids in the building of proteins. The grammar of translation is fairly discrete, with few exceptions to a simple set of substitution rules. Therefore, given a specific mRNA, one is likely to know the complete sequence of the protein that results from translation. The point of origination for the protein is easily identifiable, as are the codons- or the 'words' in the mRNA that specify amino acids- along with the sequences that indicate the termination of the protein. Even further, once the sequence for a protein is known, much can be derived about its function and role in the cell from common motifs found in the amino acid sequence.

But what is the language of transcription? Here the grammar is far less clear. The information that directs the decision to express a gene is the accumulation of extra-cellular and intra-cellular signals that culminate in the formation of transcriptional machinery on what is known as the promoter region of a gene. How and why this event arises is the subject of the study of transcriptional control. This involves the marriage of protein studies, molecular biology, and the relatively new field of bioinformatics, which applies mathematical tools to discern information found in the genomic sequence itself. To date, no single approach to the study of these regulatory sequences has been sufficient to form a predictive model. This new age of genomics will drive the collaboration of researchers from many different disciplines and necessitate new tactics in the study of the very old problem of genetics.

ARE YOU MAN OR MOUSE?

> "If you give me the coding sequences of a chimp and the regulatory sequences of a mouse; I'll give you a mouse."—unknown geneticist (via Robert Anholt, North Carolina State University)

As the public and private efforts to sequence the human genome came into force, ancillary projects also began to sequence the genomes of several bacteria, a yeast and the invertebrates: *Caenorhabditis elegans* (a nematode) and the fruitfly *Drosophila melanogaster*. These have long served as model organisms in biological research, and the complete of the sequence of the fruitfly genome in 2000 was from the most complex organism to date (Adams et al., 2000). Drosophila have been the workhorses of genetics research for decades, and having the complete sequence of the genome harkens a dramatic change in how science will progress. In particular, the volume of material available has increased dramatically, along with new avenues of investigation that would otherwise not be possible. Also whole genome comparisons are now possible both between distantly related species, such as yeast-nematode-fruitfly studies; and between closer neighbors such as the mouse and human genomes (Miller, 2001). These have yielded data and relationships between species that are a departure from what was anticipated (Shabalina et al., 2004).

One of the early surprises that came from the parallel sequencing of the genomes from several organisms is that apparent organismal complexity does not necessarily arise from a greater number of genes in the genome. From the yeast, *Saccharomyces cerevisiae*, to humans, there appears to be a common "tool kit" of about 3000 genes, which likely govern basic cellular housekeeping chores (Hodgkin, 2000). This is not a surprising result; however the relative number of genes found in each genome is. The

~13,600 genes of the Drosophila genome that were identified through research data, alignments to EST (Expressed Sequence Tag) databases, and assorted gene finder programs is only slightly more than twice that of the yeast and less than the ~19,000 reported genes of nematode. The plant model system *Arabidopsis thaliana* has at least twice as many coding regions as the fly (Rubin et al., 2000). This means that if we only consider the number of genes in these organisms, they are ranked in the opposite order than what would be expected in terms of anatomical complexity and cellular diversity. One explanation for the disjunction between the complexity of an organism and the number of protein-coding genes that it has, is that plants often utilize multiple gene copies as opposed to the complex alternate splicing found in the fruitfly; and the worm genome has undergone duplications of many large segments leading to multiple copies of genes that may be redundant (Gu et al., 2002).

Both the human and murine genomes are the same size: approximately 3.1 billion base pairs. However, while it was estimated that the human genome would contain some 100,000 genes based on the average size of the known genes and the overall size of the genome, only a little more than 30,000 have been annotated by computation or direct evidence in both genomes (Venter et al., 2001). A similar situation is true for the fruitfly genome. Of a total of approximately 180 Mb (million bases), 120 Mb is euchromatin, that region which actively undergoes transcription, while the remaining is hetero-chromatin, which consists of long (many megabases) tandem repeats of short simple sequence elements (Adams et al., 2000). These include centric and telomeric regions of the chromosomes and other sites that are likely involved in the husbandry of the chromosomes. The euchromatin in flies has far more material than that necessary for the 13,666 known coding sequences (Misra et al., 2002). It is this remainder of the genome

that is the target for annotations using the techniques presented here.

Perhaps even more surprising is that there is a nearly one to one correspondence between the known human genes and those found in the mouse. Nearly every human gene has a counterpart in the mouse (Modrek e Lee, 2003). Fewer than a score of genes unique to either genome have been found and there is a high degree of synteny, or overall organization of the chromosomes (Wiehe et al., 2000). This raises the question of wherein lay the difference between these mammals? If the number of coding regions and over-all size of the genomes are nearly identical, how is it that the products are so dramatically different in form and function? The answer may be found not in the genes themselves, but in the regulatory sequences.

TYPES OF GENOMIC SEQUENCE

A genome can be divided into three general portions: coding, those sequences specifically encoding the order of amino acids in the protein make up of cells; structural, sequences responsible for the husbandry of the chromosomes, such as telomeres, centromeres and long tandem repeats that are likely candidates as attachment sites for the chromatin matrix; and the remainder, the intergenic regions. There are also transcribed genes that do not code for proteins, but rather RNA molecules that are involved in many cellular processes including translation, RNA splicing, telomere extension, and possibly gene silencing. The coding portion of the genome comprises less than 5% of many genomes, including mammalian and Drosophilid (Venter et al., 2001; Gregory et al., 2002; Adams et al., 2000). If the proportion dedicated to the structural heterochromatin is allowed an estimated 40%, then the more than half of the genome that remains may be considered as intergenic region (Biessmann et al., 2005). That is, the sequences

“between” genes as they have been traditionally annotated. This can be further divided into untranslated regions (3’ and 5’ UTRs), which are the parts of the transcripts that do not directly code for protein sequence. These distal portions of the resulting transcripts may be responsible for the management of mRNA. Events such as sequestering and determination of the rate and duration of translation may be encoded in these sequences (Minshall e Standart, 2004; Lorenzini e Scheffler, 1997).

The remainder of these intergenic regions were previously, rather arrogantly, dismissed, as "junk DNA" in popular and peer-reviewed journals (Wong et al., 2000). Recent research indicates that a major fraction of this material is dedicated to transcriptional control (Kraft e Horvath, 2003). It may be here, in these regulatory sequences, that the information which dictates how the predominantly similar genes in mice and humans can lead to such obviously different outcomes.

ONE PERSON’S JUNK IS ANOTHER’S TREASURE

In metazoans, thousands of genes must be differentially expressed depending upon the tissue type, the developmental stage, and environmental factors present. Eukaryotic gene expression is a coordinated interaction between proteins (the trans-acting component), cis-regulatory elements in the DNA along with extracellular and intercellular signaling pathways. The combination of these factors leads to the assembly of the initiation complex, including RNA polymerases, onto the sequences that comprise the core promoter. It is only after this complex has formed that transcription may take place. With some exceptions, such as RNA editing. transcription and translation will reliably follow once the transcription initiation has occurred. This is illustrated by the ability of a host organism to produce a functional protein from a transgene derived from a

completely different organism. Regulatory sequences, on the other hand, more typically *are* species specific. Some basal promoters will drive expression in foreign species, but the particular circumstances that lead to regulation of transcription are lost in transgenic sequences. This reinforces the notion that it may largely be the regulatory sequences that distinguish organisms, rather than the coding sequences, which to some degree, are interchangeable.

The Core Promoter

The core promoter is typically described as a region approximately 80-100 base pairs (bp) surrounding the initiation site. In particular, there are specific sequences that serve as recognition sites for 'basal' transcription factors such as TFIIB and TBP. Core promoter elements include motifs called BRE, TATA, Inr and DPE (Lemon e Tjian, 2000).

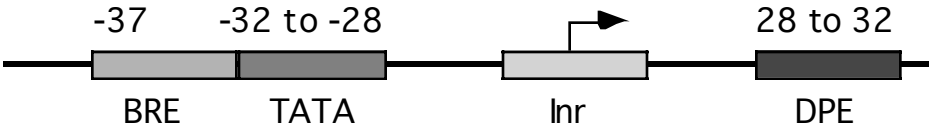


Figure 1: Relative Positions of Core Promoter Elements

Some of the general motifs found in core promoters. They differ by position and component of transcription complex bound. The TATA box, which is rich in T and A bases, is found around 30 bases upstream of the +1 site (start of transcription- arrow). The BRE or TFIIB recognition motif is an extension of a subset of the TATA boxes known. The Initiator (Inr) surrounds the +1 site and is found with or without TATA elements. The DPE (Downstream Promoter Element) is found symmetrically opposite to the TATA element, but more commonly in TATA-less promoters. (after (Butler e Kadonaga, 2002))

Other types of core promoter sequences have been identified in Drosophila using a motif recognition algorithm, MEME. This program was applied to all the transcription

start sites (TSS) for most of the genes in one arm of chromosome 2 (Ohler et al., 2002). The authors first aligned the entire chromosome arm sequence to a collection of full-length (cap-trapped) cDNA library sequences. This allowed them to identify at least one putative promoter, or TSS, for each gene. The 120 bp region around the transcription start sites for these genes were compared and 10 distinct classes of promoter elements, including the four above, were found. The study is important in that it takes a blind approach to finding the core promoters— that is, MEME recognizes motifs based on the non-random assortment (or motifs) of small sequence fragments. MEME relies on a so-called background file which is an assessment of the general randomness of the entire target sequence. That is, it calculates the rate of occurrences of not only each base (A,T,C,G), but also how often groupings of bases occur (such as doublets: AC, GC; and triplets: AAA, GGG, or ACT). This background file is a critical step for increasing the efficacy of the MEME search algorithm. MEME is not pattern-matching to a set of previously described motifs, it is identifying a commonality between these specific sites based on their sequence pattern relative to the surrounding sequences (Grundy et al., 1997). By tying this information to biological data that also identifies putative promoter sites (such as RACE and EST matches), the authors were able to describe novel and variant core promoter elements. The intersection of the two data sets was smaller than expected, implying that each technique will miss genuine promoter sites. They note that this is important because it demonstrates that a simple search for the 'typical' motif will often fail to find a core promoter for a particular gene, even if the search is focused upon the upstream sequences for that gene. Ohler and others have found that less than 45% of the 13,000 Drosophilid genes have a recognizable TATA element and fewer have a DPE in the expected location (Arnosti, 2003).

Another critical point of their data is that it draws to light the necessity of knowing the site(s) of transcription initiation before an analysis of what types of promoter sequences exist and the frequency with which they occur can take place. The data from the cap-selected cDNA library may be incomplete, causing them to search for core promoter motifs in the wrong place. Without the TSS, the current algorithms fail to distinguish basic core promoters. Additionally, there may be more than one TSS for a given gene and completely non-standard promoter arrangements have been found that complicate the search (Lee et al., 2005). This is one of the themes of this work— that no single approach is sufficient to define the cis-elements involved the transcription of a gene. Information from molecular and transient expression studies is necessary to reliably map the promoters and further, the regulatory elements that affect them.

In addition to the sampling of core promoter elements described above, recent data have revealed new promoter sequences in Drosophila that operate independently of the classic TATA or INR-DPE type promoters. One such example is the TC box, which is found further 5' than the TATA box (at about -77). Whereas TBP (TATA binding protein) has been ubiquitously found in the pre-initiation complex, TC boxes are instead bound by TRF-1 (TBP related factor) which is expressed in developing neural tissue (Holmes e Tjian, 2000).

There is little doubt that there will be many more elements and types of promoter-transcription factor interactions discovered in the future. It has also become apparent that as conserved as the process of transcription initiation seems to be, there are also many species-specific factors and cofactors that have been found, which add to the complexity of the task of finding promoters and the enhancers that regulate them. Functional studies that employ transgenes are a primary path used in this research, but the addition of

bioinformatics will increase the efficiency of this type of work greatly.

Cis-Regulatory Elements

In addition to the core promoters, there are other known elements in the regulatory sequences affecting the transcription process. These cis-acting elements include proximal promoter elements, enhancers, silencers, and insulators or 'boundary' elements (Gerasimova e Corces, 2001). These are the players in a combinatorial scheme of transcription regulation that allows for an extremely complex set of circumstances to create an intricate array of gene products (Smale, 2001). In order for this to occur, remote enhancer elements likely recruit factors that lead to chromatin remodeling (Ho et al., 2002). This, in turn, allows for the increased likelihood that transcription factors will bind to their respective enhancer and repressor sites proximal to the promoter (Elefant et al., 2000). It is at this point that some of the specificity might occur. Depending upon the order and type of remodeling that takes place, particular binding sites can be exposed or hidden to particular transcription factors (Gregory e Horz, 1998). There is a concert of acetylase, methylase and phosphorylase activity that is directed by as yet unknown cues, which modify the histones that bind and package genomic material and dictate whether other DNA binding factors are permitted access to specific regions. It is this chromatin remodeling, along with accumulated effects of the transcription factors recruiting the initiation machinery to the promoter, which will determine if transcription is to proceed.

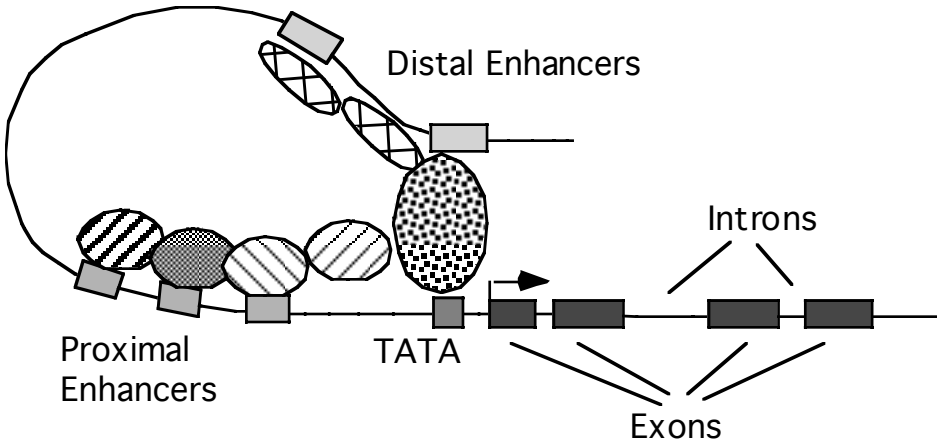


Figure 2: Map of Regulatory Elements

This is a generalized schematic of the assorted elements involved in transcriptional regulation. The dark grey boxes delineate exons. The lighter grey boxes represent the other DNA elements contributing to initiation. Additionally, some elements may be downstream of the coding region, within introns, or may even exist beyond another gene. The collection of transcription factors generates an environment conducive to the formation of the initiation complex. In some cases the entirety of the regulatory elements are within a few hundred bases of the transcription start site, while in the other extreme they have been located several hundred thousand bases away, such as with the bithorax complex (Bate e Martinez, 1993; Lewis, 1998). It stands to reason that the more complex the expression pattern, the more extensive the regulatory region might be. Some genes are constitutively expressed, while others are only needed at specific moments in development. Several studies have shown that for many genes a kilobase of sequence typically contains the complement of regulatory cues (Arnosti, 2003).

Cis-elements are typically referred to as enhancers, but as the information about these sequences and their function accumulates more discrete descriptions will be necessary. In general, enhancers are binding sites (or collections of sites) for transcription factors. These factors can either promote or inhibit the onset of transcription either by modification of the chromatin structure or involvement in the recruitment of the initiation complex (Blackwood e Kadonaga, 1998). The specificity of one factor for a designated site is difficult to define, *in vitro* studies have often disagreed

with *in vivo* evidence, but there is building repertoire of sequence-transcription factor matches. The regions defined as enhancers will often contain multiple binding sites for a group of factors working either cooperatively or in a competitive manner. Enhancers can function as integrators of different expression information. Often information regarding tissue and developmental specificity can exist in separable cassettes that combine into spatial and temporal patterns of control (Brenner et al., 1996; Small et al., 1992).

Conceptually enhancers are passive sensors that detect regulatory information. In a physical sense they are landmarks that the transcription factors use to arrange themselves. Therefore, the information such as the instructions for a specific pattern of a gene expression is dependent also upon the suite of transcription factors present. In the case of a complexly regulated gene— one that is expressed under many specific circumstances in time and space— the cis-regulatory elements associated with that gene may be the equivalent of a simple on/off switch. The complexity may arise in the regulation of the transcription factors themselves; not in the gene's control region. This would explain how a complexly regulated trans-acting element can lead to complex regulation of a gene with a relatively simple regulatory region. Another aspect of enhancers is that they are often position-independent in that an arbitrary re-arrangement of elements has been shown to recapitulate the same expression pattern of a transgene. They may be located upstream or downstream of the core promoter and often within intronic sequences between coding exons (Merika e Thanos, 2001).

There are many more complexities to add to the description of enhancer elements including shared regulatory sequences, promoter competition, and discrete enhancer-promoter specificity regardless of the intervening sequences (Butler e Kadonaga, 2001). Another issue mentioned in the case of the *giant* transcription factor is the clustering

nature of some recognition sites. This may be a case of a democratic accounting in transcriptional regulation or an adaptation to spatial constraints in forming a cooperative monolithic transcription factor complex. Another consideration in binding site clustering is that it may simply serve to increase the likelihood of binding by increasing the density of the target site for a dilute factor.

All of these disparate aspects of regulation require an armada of tools to approach the different problems presented in the varied regulation of genes. Reiterating the argument that while coding sequences are typically interchangeable between species and that evolutionarily distant organisms share much coding sequence but relatively little regulatory sequence, it follows that information regarding the distinctive form and physiology of a species lies in the transcriptional control regions rather than the coding sequences.

Chapter 2: Promoter Bashing

Previous Studies of Cis-regulatory elements

Functional analysis of promoters

Early studies of the elements involved in gene regulation relied on cell-culture lines but these are fraught with cell-specific discrepancies. In studies that sought to define the redundancy of the core promoter TATA and Inr elements, *in vitro* and cell culture assays suggested either element was sufficient for expression, but later *in vivo* assessment demonstrated each component was necessary and not interchangeable (Wu et al., 2001). A genetic study of the Antennapedia complex genes showed that the group of more than 12 genes specify segmental identity of the developing embryo. These segment-specific patterns are established through a collection of pathways including autogenous control, cross-regulatory interactions between homeotic proteins, and trans-activators— including the gene: *Sex combs reduced* (*Scr*) (Paro, 1990). In a study of how *Scr* is itself regulated, the surrounding 100 Kb of genomic DNA was restriction-digested into 3-10 Kb pieces that were inserted into an enhancer trap vector that had a minimal HSP70 promoter driving expression of the LacZ reporter gene. The fragments used spanned from within the coding region past an upstream gene on an opposite strand (*ftz*) and further to yet another gene: *antp*. The expression patterns of these constructs were then compared to an *in situ* analysis of *Scr*. Several interesting patterns were revealed, but there were difficulties in interpreting whether the elements responsible for expression patterns similar to the endogenous gene were affecting *Scr* or neighboring genes that also express in similar tissues. Some elements were found to be 25 Kb away, beyond an intermediary gene. The study also found fragments that directed expression patterns

similar to those found with mutations in neighboring genes in the complex, suggesting a competition for these enhancer elements between the promoters for the different genes (Gindhart et al., 1995).

To date, the suite of factors that dictate the early development of Drosophila embryos has been extensively studied. However, even when armed with distinct binding sequences for the *giant* gene product, a great deal of bias had to be introduced into a genome-wide search for these sites near the genes known to be regulated by *giant*. There was a large array of functional evidence to support the findings, but the search target needed to be enriched for non-coding sequence along with the presumption of clustering of the binding sites. After these accommodations, the authors were able to locate *giant* binding site clusters near to only one gene, *even-skipped*, of several known to be regulated by *giant* (Berman et al., 2002). Reversing this approach, they found that by searching for the binding sites for all the known regulators of *giant*, this allowed them to find a cluster 2.5 Kb upstream of *giant*. However, they are quick to point out that they had a great deal of prior knowledge that allowed to find this cluster and that the grammar of the cis-regulatory region is largely unknown— only that sites are clustered, and found in non-coding sequences. This nascent bioinformatics approach to enhancer sniffing highlights the continued need for functional analysis in discerning the role of putative enhancers. *In vivo* studies such a DNAse protection and gel-shift assays contribute when transgenes prove to be challenged by issues such as the size of the region studied (Beaty e Marquardt, 1996). These identify the minimal target sequences bound by proteins, but are plagued by the variability presented in an *in vitro* environment.

The search for conserved non-coding sequences

Of the several dozen sequenced metazoan genomes reported to have 3-5% coding for protein, and a further 40% masked out as tandem and dispersed repeats, there still remains half of the genome as non-coding (Smit, 1996). Analysis of non-coding regions has been in the realm of gene- or locus-specific studies using transgenic lines for some time. With the availability of many genomes fully sequenced, there is now a tool for locating functionally conserved elements by comparing these regions to one another. In particular, Loots, et al, compared nearly a megabase of the mouse genome to the orthologous region in the human genome to identify conserved non-coding blocks. The region of interest is a locus the codes for 5 human interleukin genes along with as many as 18 others. They were able to show that the deletion of one of the larger conserved blocks reduced the production of three interleukin genes, one of which is 120 Kb distant (Loots et al., 2000).

Of particular import to this work, whole genome comparisons between multiple mammalian sequences showed that non-coding RNAs and many previously described regulatory sequences are *more* conserved than coding sequences (Dermitzakis et al., 2003b). Because so many earlier studies had focused on only coding sequences, the amount of the genome that is under selective constraint has been estimated to be less than 3%. Here, Dermitzakis *et al* report an amount nearly twice that with the majority of these conserved blocks falling outside of coding regions. In a second publication, they measured the rate of polymorphisms of well-described transcription factor binding sites in several species of Drosophila and found that these sequences exhibit patterns of variation consistent with functional constraint. This was surprising in light of the belief

that regulatory sequences undergo high turnover between species (Dermitzakis *et al*, 2003a).

Summary

The suite of tools available for the task of ferreting out the nature and language of regulatory sequences in promoter regions is expanding daily. Whole genome sequencing projects have brought to light relationships within the structure of the control regions as well as the evolutionary changes and conservation of sequences between species. While *in vitro* and *in vivo* analyses cannot be replaced, the application of new algorithms and models are yielding new insights in the search for the determinants of transcriptional control

The particular case of slowpoke

Having outlined some of the generalities of work in the field of transcriptional control, here is a specific case in which a promoter region has been dissected into sequences that control spatial, temporal, and conditional expression of a gene under especially complex regulation. A mutation in a calcium-activated potassium channel in Drosophila causes a "sticky-feet" behavioral phenotype. This channel is the homolog to the mammalian BK channel. *In situ* analysis and separation of the promoters mapped with 5' RACE revealed that *slowpoke* is expressed in the nervous system, musculature, and in parts of the gut and tracheal systems. The expression in these different tissues was found to be driven by four distinct promoters and the regions in the control region responsible for this expression was delineated by a series of expression constructs and deletion analysis (Brenner e Atkinson, 1996; Brenner et al., 1996). A fifth promoter was later identified further upstream from the most 5' TSS, through sequence comparison

with another species of Drosophila. A more sensitive RACE reaction was performed, guided by the conservation of sequences between the species, yielding another neuronal promoter. This highlights the utility of sequence analysis to augment molecular work in defining promoter regions (Bohm et al., 2000).

The *slowpoke* promoter is particularly complex with five distinct (and possibly more) transcription start sites and two different translation start codons in the muscle and neuronal transcripts. There were also regions in the regulatory sequences that dictate early embryonic and larval expression and others specific to adult stages. Further work revealed that other conserved elements in the area of the muscle promoter activated expression in distinct muscle groups and tracheal cells of the fly (Chang et al., 2000). More recent work has shown that *slowpoke* expression is elevated upon exposure to alcohol, and that the neural promoter is specifically necessary for this response (Ghezzi et al., 2004). Chromatin Immuno-Precipitation (ChIP) assays have localized some of the sequences within the promoter that are associated with altered histones in an apparent case of chromatin remodeling prior to gene expression. There is a distinct shift in which portions of the regulatory region are associated with acetylated histones following alcohol treatment. This combined set of results illustrates a case where regulatory elements could be identified through diverse means that dictate expression spatially, temporally, and in response to an environmental insult. It is certainly true that not all promoters and their associated regulatory elements will be this complex, but it may be that the diverse roles that ion channels play in the physiology of an organism requires such elegant control.

Ion channel gene regulation

The example of the *slowpoke* transcriptional control study demonstrates that in

some cases simply scanning through fragments of upstream DNA in a functional assay will fail without knowledge of the extent of the UTR exons and conserved non-coding regions. Many of the efforts to de-cypher the control elements for a gene in the last five years have involved genes that play a role in early development. These genes, such as the *Hox* gene cluster, establish segment identity and appear to have discrete and ordered interactions during development. On the other hand, ion channels such as *Slowpoke*, are expressed ubiquitously— virtually every cell has some set of channels— and more interestingly, are a critical component in highly plastic tissues such as neurons and cells of the immune system. It stands to reason that a gene which is expressed in a great diversity of tissue types and that plays a role in altering the excitability of those tissues over time should be under complex transcriptional control.

Complex behaviors are ultimately underlain by the nervous system. This, in turn, is the sum of the individual states of all the neurons and their interactions. Each neuron's resting and excited state is dictated by the action of the suite of ion channels extant in the cell. It stands to reason then, that ion channels are at the root of behavior along with many other complex physiological aspects. Lastly, the control of the expression of these channels in a cell lies at the foundation of the emergent properties of behavior, learning and memory.

Ion channel genes, particularly potassium channel genes in Drosophila, are a diverse family with scores of members. They control the flow of ions across cellular membranes, which in turn, establishes the electrical climate of a cell. They also serve in intracellular signaling processes and transport. Families of ion channels have been described based on their electrical properties or their sequence homology. While *slowpoke* is a calcium-activated potassium channel, *Shaker* is a voltage-gated channel

that conducts the A current in cells. The two are major components in the transient potassium flow that re-establishes the resting potential of a cell membrane after depolarization. There are also at least three other members of a so-called "Shaker cognate" family of ion-channels: Shab, Shal, and Shaw. Recently some have begun to assign a different, less confusing, nomenclature adopted in the murine system: Kv1 (*Shaker*); Kv2 (*Shab*); Kv3 (*Shaw*); and Kv4 (*Shal*). This numerical system will be used here.

Originally classed as a family of genes by sequence similarity to the pore region of the *Shaker* protein, the cognates have diverged in function over time (Butler et al., 1990). Kv4 conducts a small outward potassium current similar to *Shaker* and s*lowpoke*. Kv2 and Kv3 are delayed rectifiers, remaining open as long as the membrane potential is greater than their threshold point. While all of these channels are found in neural tissue, there is some evidence of different sub-cellular localizations and cell-type specificities between them (Baro et al., 1994; Baro et al., 1996; Schneider et al., 2000). The human homologues of these and other potassium channels have been implicated in several disease states including the Long-QT and 'Jervell & Lange-Nielson' syndromes; episodic ataxia, deafness. and heart failure (Pfeufer et al., 2005; Winslow et al., 1998). In particular, a down-regulation of an otherwise fully-functional delayed rectifier potassium channel in humans has been targeted as a cause of transient arrhythmogenesis (ventricular tachycardia) (Nabauer e Kaab, 1998; Pogwizd et al., 2001). Knowing how this change in transcriptional control occurs, and being able to recognize a genetic pre-disposition for this disease by understanding more about the regulatory sequences affecting these genes, is essential for the development of pharmacological strategies in the treatment of sudden heart failure— a leading cause of death in the US today.

Here then, is where the questions begin- how are channel genes regulated? Is there an interaction between the different genes with respect to their transcriptional control? Are there recognizable common elements in their promoter regions? As a step toward answering these questions, the control regions must be characterized.

RESULTS and DISCUSSION

Veni, vidi, vici. —Julius Caesar. Suetonius, Lives of the Caesars.

in vitro, in voltro, in vivo. —*inveterate* graduate student.

The following three chapters deal with different approaches to the search for regulatory sequences. In each chapter, the results and discussion of the data are combined. The results are discussed as they are presented. The first section, *in vitro*, describes the data obtained from mapping the complete cDNA back to the genomic material, and the cloning of the same sequences from other species. The second section, *in voltro* (*in silico*), is a survey of the sequence analysis performed to annotate target sequences that were used in a functional analysis using transgenic flies, which is the final section: *in vivo*.

Chapter 3: *in vitro*

FROM CODING SEQUENCE TO RACE TO GENOMIC MAP

In order to describe the control elements for a gene, it is first necessary to define the reasonable limits of the region involved. Traditionally, a gene referred to the set of genomic sequences that coded for a complete protein product. As more expressed sequence accumulated, the annotation of genes expanded to include the sequences of the complete mRNA including untranslated exons at both ends of the transcript. It was also recognized that in another 100 bp or so of sequence upstream of the first exon lay the promoter for that gene. However, a gene could be described as the set of sequences necessary for the coding of the protein, the complete mRNA, and that contain the information regarding expression of that gene— which is a critical component of the

phenotype of a gene. For our purposes here, it is assumed that these regulatory elements lay within a few kilobases of the most upstream promoter (or 5' most exon). So it was first necessary to delineate the extent of the region containing all the exons for the genes in this study, then the same for neighboring genes to cordon off a reasonable limit of material to be cloned.

At the outset of the project, the genome had yet to be sequenced and only coding fragments of the cDNAs for the four members of the *Shaker* family were extant. These cDNAs were used as a starting point for the *in vitro* portion of the project— obtaining full-length messenger RNA. They were sequenced and primers were designed to perform a 5' RACE. Rapid amplification of cDNA ends, or RACE, is a technique for cloning the full-length transcript of a gene, including the untranslated material at the 5' end of the message. The products of the RACE reactions were then mapped against large genomic clones to determine the locations of the promoters for the genes. This data, married with sequence analysis, would serve as guide for designing genomic clones and expression vectors. Only two of the genes were considered candidates for sub-cloning into transformation vectors, due to size constraints. These constructs are then used to generate transformant lines for a functional analysis of the control region. Of these two parallel efforts, one came to fruition and is presented here.

cDNA Cloning

The work on the three *Shaker* cognates: Kv2, Kv4, and Kv3, is described. cDNA clones are obtained from library screens and via reverse transcription reactions of total RNA. These clones were then sequenced and used to design primers for subsequent RACE reactions (see Materials and Methods).

Kv2

The known current cDNA clone was mapped against a genomic clone to locate exons and a possible 5' UTR. At the outset, only a small (331 bp) fragment of Kv2 was available (kindly donated by the Thummel lab). This was radio-labeled with ^{32}P and used to probe a cDNA library. Of 106 positive colonies, 84 were picked, re-plated and screened with the same probe under more stringent conditions (higher temperature and more blocking agent). Nine clones were selected and sequenced, yielding four that matched the known sequence of the Kv2 (Shab) protein. These four clones revealed a splice variant in which exon 4 is excluded, with a common transcription start site just a few bases upstream of the first putative methionine codon. The other five clones were either other *Shaker* cognates or genes not previously described.

The sequence of this clone was mapped against the genomic sequence to delineate exon boundaries and untranslated regions. The unique 5' ends of the clones serve as landmarks for the locations of the promoters and the organization of the genomic region. However, all four clones began with the four-base leader before an ATG coding for the putative initial codon. We could not rely solely on the position of the annotated 5'-most exon because it was not uncommon at this time to report only the minimal cDNA sequence that includes just the coding region. Untranslated sequences were less useful in

transgenic studies and increased the size of the construct. It was later shown that including the first intron (and thus, the next 5' UTR exon) could often increase the efficiency of expression of the transgene, and so a greater effort to locate the 5' ends of cDNAs progressed.

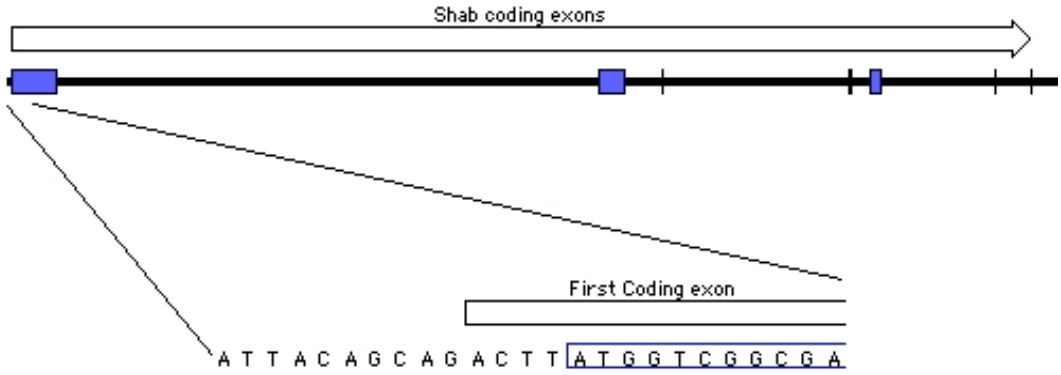


Figure 3: Map of Kv2 (Shab) Coding Region

The seven known coding exons are shown as blue boxes (or bars in the case of small exons). The inset is an enlargement the 5' end of the first coding exon showing the four base UTR upstream of the first methionine codon.

The seven coding exons were mapped to a 29 Kb region of the P1 clone. The published transcription start site agreed with the four cDNA clones pulled from the library. However, it was suspected that this was not the most 5' start site. Often cDNA library clones can be biased to terminate at a particular restriction enzyme cut site. The full-length cDNA was needed, including 5' UTR exons in order to be able to map the true transcription start site. To extend the cDNA clone, a RACE reaction would be needed.

Kv4

Primers to the reported Kv4 coding sequence were designed and a cDNA was cloned from total RNA. This was sequenced and three single base differences from that reported in Flybase were confirmed. This was important in designing primers for the

subsequent RACE reaction.

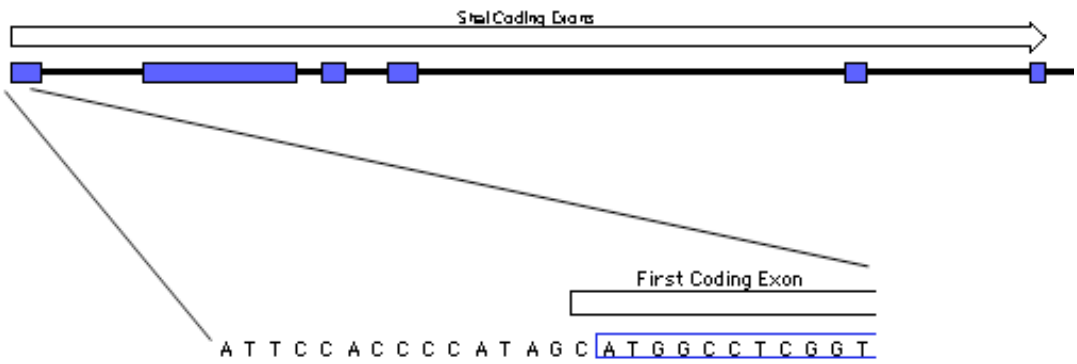


Figure 4: Map of Kv4 (Shal) Coding Region

The six known coding exons are shown as blue boxes. The inset is an enlargement the 5' end of the first coding exon showing the single cytosine base upstream of the first methionine codon.

The six exons comprising the coding region span less than 7 Kb. Only a single base is annotated upstream of the start codon. While it is possible to have transcription and translation begin at or near the same nucleotide, we considered this to be unlikely and decided to more carefully map the 5' end of the mRNA. Several other ion channel genes have been described with 5' UTRs of a couple hundred base pairs or more. Additionally, reviews of the Drosophila Genome Project annotations using direct evidence and merging of Genescan (a computational predictor of exons) assessments, typically added more than 80 bp to the UTR's of genes that had no UTR's (Misra et al., 2002). Many reactions later we found that all clones still had the same single base before the start codon. We used RACE along with sequence analysis to extend to cDNA and locate another 5' UTR exon far upstream.

Kv3

Primers to the reported full-length cDNA for Kv3 (Shaw) were used to amplify the transcription products. This was cloned and sequenced. Alignment of the cDNA sequences for the Shaker cognates showed that all three cDNAs contained reasonably complete coding regions. This was indicated by the inclusion of a complete T1 domain necessary for the assembly of the subunits into a tetramer, which is found at the N-terminus of all the *Shaker* cognates.

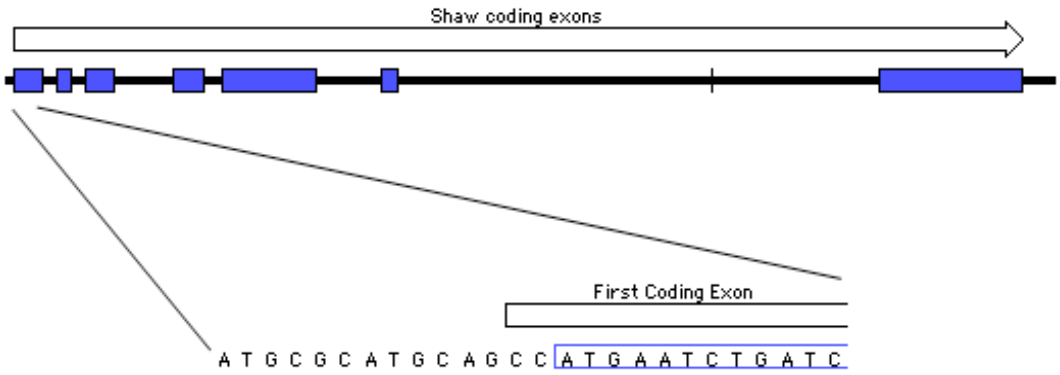


Figure 5: Map of Kv3 (Shaw) Coding Region

The seven known coding exons are shown as blue boxes. The small bar is an eighth exon that exists in some clones. The inset is an enlargement the 5' end of the first coding exon showing two cytosine bases upstream of the first methionine codon

These cDNA clones served as a steppingstone to cloning the entire control regions of these genes. The 5' end sequences were used to design primers for a RACE reaction. Comparison of the exon structure showed a similarity in the first exon with only a few bases of UTR before the start of the coding sequence. The control regions for the *slowpoke* and *Shaker* potassium channel genes have been shown to be quite complex,

with multiple 5' UTR exons and TSS's. This led us to suspect that each gene in the *Shaker* cognate family could easily have more exons, and promoters, upstream.

RACE

> An adage that holds true in life as well as in science is that the intrinsic worth of an item is often proportional to the difficulty involved in obtaining it. Although difficult to achieve, isolation and analysis of the 5' end of a cDNA provides extremely valuable and essential information.
> -- from Ambion, a company dedicated to products pertaining to RNA.

In order to map the transcription start site of the *Shaker* cognate genes, 5' RACE was performed. The resulting clones were sequenced and mapped against the genomic region. There was more than one product for each gene, signifying that either more than one transcriptional start site existed, or that there was alternative splicing in the 5' UTR. In the case of multiple transcription start sites, this would indicate multiple promoters controlling expression of these genes.

The full-length cDNAs were obtained for each member of the Shaker cognates: Kv2, Kv3, Kv4, and a new candidate member, Kv3.2. Each is shown to have two transcription start sites and 5' UTR exons upstream of current annotations in the databases. Kv2 has a large intronic region spanning over 25 Kb above the coding sequence; Kv3.2 and Kv4 have a similar intron over 19 Kb in size; while Kv3's intergenic region is less than 6.5 Kb. In choosing a region to clone for functional analysis, several factors come into play. Among these are size and complexity. The inclusion of elements such as exons and promoters from other genes can complicate the analysis. For these reasons Kv2 was excluded from the functional analysis, and Kv4 and Kv3 were attempted.

A RACE reaction is a critical bridge linking information from a shotgun genome sequence and transcription processes in cells. Typically, cDNA clones are incomplete and comprised of the central portion of the coding region that was cloned utilizing

conserved sequences. Commonly, a cDNA clone that included an upstream stop site in-frame with a methionine codon was considered complete. However, the upstream non-coding sequence of the resulting mRNA may be quite extensive and reveal pre- and post-transcriptional regulatory elements. For this reason, finding the full-length message is important. Currently, not enough is known about transcription initiation to recognize a start, or +1, site in the upstream sequences of genes. We therefore must rely on mapping the complete cDNA back to the genomic to isolate promoter sequences.

Such an important step currently relies on two technologies: a "template switch" SMART method and T7 RNA Ligase-mediated RACE (RLM-RACE). The first uses a modified MMLV reverse transcriptase that exhibits terminal transferase activity upon reaching the end of the transcript. This appends a string of cytosines at the terminus of the first strand synthesis (Zhu et al., 2001). These, in turn, anneal to an anchor sequence that has a poly-G 3' end. During first strand synthesis, the reverse transcriptase switches from the RNA template to the newly annealed DNA anchor. The second technology used was RLM-RACE. First utilized in the lab by Rudi Bohm to isolate message from a remote promoter in the *slowpoke* gene, an RNA anchor is ligated directly to the 5' terminus of the message. Additionally, the RNA has been culled of incomplete messages (those that have either degraded or that have not finished splicing). We found that the 'strand switch' technique was often confounded by complex secondary RNA structure and that the necessary mutant reverse transcriptase lacked the processivity for longer messages and extended UTR's. We therefore worked to optimize the RLM-RACE by altering many variables such as reaction temperatures, the addition of adjuvents, and primers designed against exons predicted with sequence analysis.

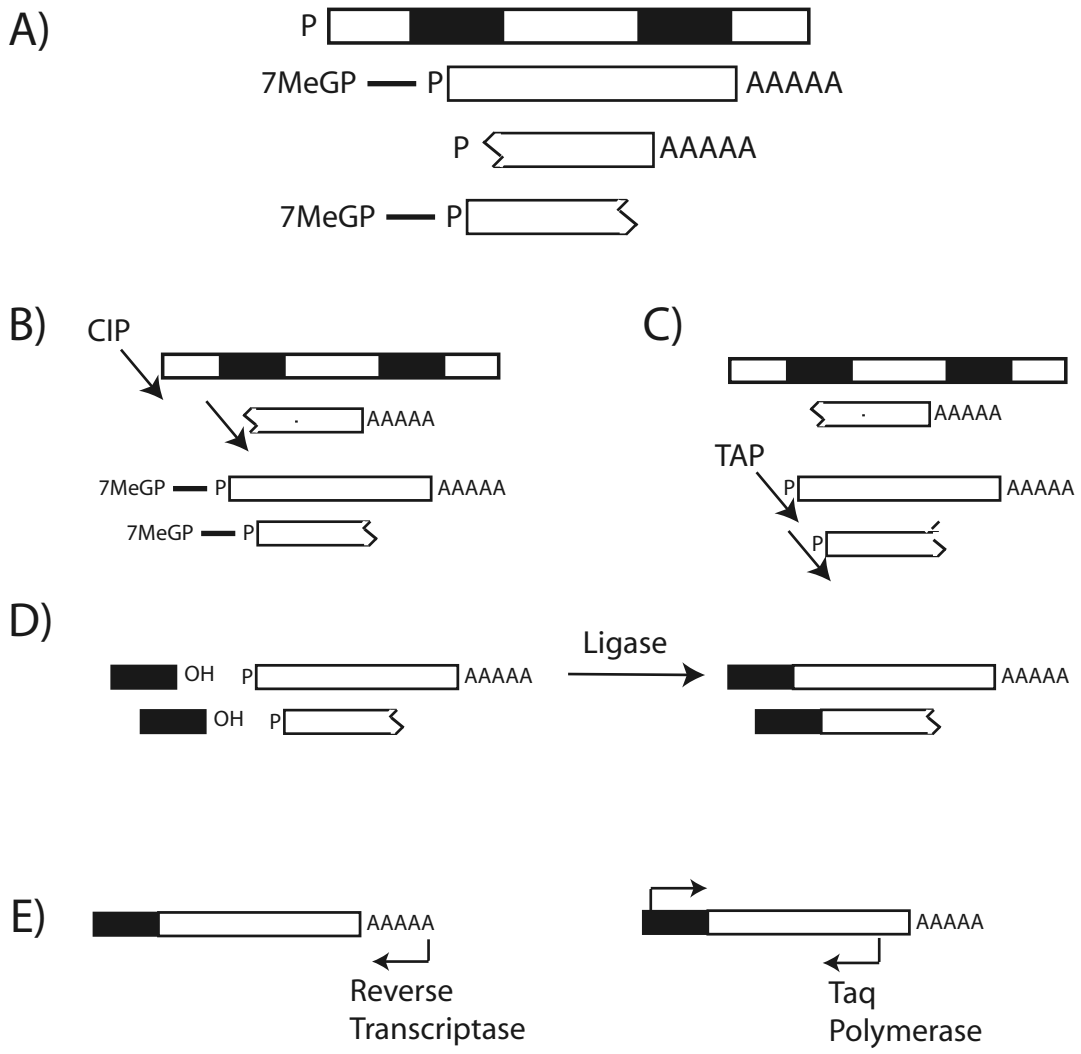


Figure 6: RNA Ligase Mediated 5' RACE

RLM-RACE increases the likelihood of obtaining a full-length cDNA. A typical pool of total RNA is represented in (A). This includes unspliced message (introns are black boxes) and degraded messages with no 5' cap or poly-adenylated tail. In (B) Calf Intestinal Phosphatase removes the exposed 5' phosphate from the incomplete message and then in (C), Tobacco Acid Phosphatase removes the 5' Cap, leaving a phosphate for the subsequent ligation reaction with an RNA anchor. The pool is now enriched with full length messages for the RT reaction with oligo-dT primer to eliminate non-polyadenylated messages.

Kv2

We first used the SMART RACE "template switch" method for Kv2, however this produced no new information. Therefore, an RLM-RACE reaction was run along with a high-temperature reverse transcription reaction, to identify three new exons for Kv2.

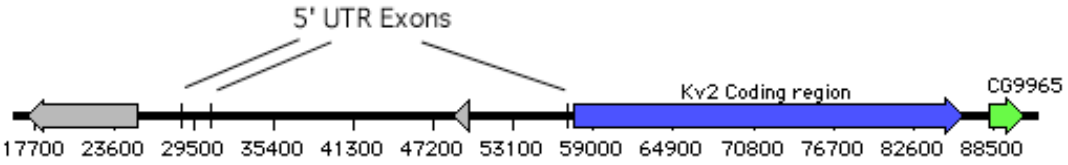


Figure 7: Map of Genomic Region Around Kv2 (Shab)

Three additional exons were identified in the 5' UTR of Kv2. These are shown as small bars upstream of the coding exons. Two unique clones with different 5' start sites were found. One isoform added only the first exon upstream of the first coding exon, while the other included this exon along with two more upstream exons.

The grey arrows (one spanning bases 17,600 to 28,000 and another at 50,000) are annotated genes on the complementary strand. The green arrow is the next gene downstream of Kv2 on the same strand. The numbers represent the relative position on the genomic BAC clone.

The 5' RACE reaction of the Kv2 (Shab) cDNA revealed that the most upstream exons lay more than 25 Kb away from the coding exons. Additionally, a portion of this upstream region included a match to a published cDNA fragment found in the Drosophila EST (Expressed Sequence Tag) database. However, this recently annotated gene is on the opposite strand from Kv2. Having a nested gene in the control region introduced the complication that any regulatory elements in the region may actually drive expression of this other gene rather than Kv2. This would confound dissection of the regulatory region using reporter constructs. If a deletion is made that removes a regulatory element for the target gene, the control region from the neighboring gene might be brought into

proximity such that it could lead to expression of the reporter gene in a fashion similar to the secondary gene. The same is true should any boundary elements be deleted. The size of the clone necessary for a functional analysis of the control region would exceed 38 Kb. While it is possible to produce such a large transformation vector, this would be intractable for producing deletions (Payne et al., 1999).

Lastly, since at least one gene appears to be included within the control region, this leaves room for some doubt as to whether the flanking genes on the same strand are genuine boundaries for the control region. It has long been known that for some genes, regulatory sequences might be many kilobases distant, far beyond neighboring genes. For these reasons, size and the inclusion of another gene between 5' UTR exons, further work on cloning the control region for this gene was halted in favor of the other two genes in the *Shaker* cognate family. This situation highlights the need for new tools and different approaches to promoter analysis. Perhaps, in combination with sequence analysis, inter-species comparisons, and chromatin remodeling assays, the scope of the target sequences involved might be narrowed.

Kv4

For the next member of the Shaker family, Kv4 (Shal), it proved to be far more difficult to clone a complete cDNA. Several techniques employed in the Kv2 RACE failed here. Two factors that might affect a successful reverse-transcription reaction necessary for full-length cDNA products are secondary structure in the mRNA and very long UTR's that exceed the processivity of the polymerase. We used a new reverse transcriptase, Thermoscript (Invitrogen), to run the reaction at a higher temperature, thus denaturing any secondary structure present. This allowed us to extend the length of the first coding exon to include 352 bp of UTR. Computer analysis and direct inspection of

the sequence were used to identify additional 5' exon candidates. Two different methods were employed to identify a possible upstream exon, to which a primer was designed for the PCR step in the RACE. First, canonical exon boundaries and core promoters were annotated. Along with this data, a Fickett's test for non-random nucleotide distributions was overlaid upon an alignment of sequences from four species of Drosophila (see Genomic Cloning). The coincidence of conserved sequences implying some functional constraint upon variation, and the exon boundary consensus sequences along with an upstream promoter motif, led us to believe that this was an exon in the Kv4 transcript. The resulting PCR product included this exon in a nested primer reaction.

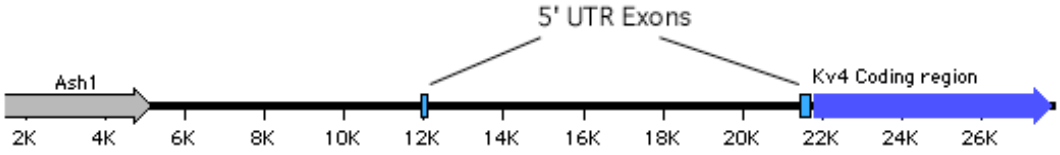


Figure 8: Map of the Genomic Region Around Kv4 (Shal)

Sequencing of the 5' RACE clone revealed that the first coding exon extends more than 350 bp further upstream of the translation start site than previously reported. An additional 5' UTR exon was also found to exist 10 Kb upstream. The coding region is in dark blue while the UTR is in light blue. The grey arrow is the next gene upstream on the same strand. The numbers represent the location on the genomic clone.

RNA specifically prepared from embryos preferentially showed a greater abundance (over ten fold, by gel image analysis with NIH-Image) of message originating from the 5' promoter than the downstream promoter. The reverse was seen in adult RNA. This may be a case of temporally specific promoters directing expression at certain life stages. One of the reasonable limits to a control region is the 3' terminus of the next gene upstream from the target gene. This next gene was annotated in Flybase as *Ash1*. However, only the coding region was described. To determine if there were more exons that would encroach upon the Kv4 regulatory region, a 3' RACE was conducted to

map the UTR exons. This expanded the extent of the Ash1 over 2 Kb downstream towards Kv4. The mapping of the Kv4 UTR exon 10 Kb upstream of the coding region and the annotation of the next upstream gene on the same strand set logical boundaries for the Kv4 control region. A complete transformation construct would be over 32 Kb. While this is just at the upper limit of a reasonable size of vector for us to inject into flies, the efficiency of such transformations can be quite low. Additionally, cloning large fragments pose difficulties in normal bacterial hosts that tend to trim plasmids in the face of an antibiotic challenge. Lastly, ensuing directed deletions with endonucleases would be difficult to perform due to the large number of restriction sites in the entire fragment. It was therefore decided to divide the region into thirds and clone the promoter fragments into an enhancer trap vector (pPTGAL) driving a Gal4-UAS reporter system. As described in the *in vivo* section, this presents other hurdles in the functional analysis of a control region, namely that the elements in the region may not be modular. They may require that all portions be present for a reasonable recapitulation of the native expression pattern of the gene. Separately, there may be no expression within the bounds of the native expression. This, in fact, appears to be the case. The constructs generated from the arbitrarily designed "thirds" yielded minimal expression.

Kv3

The last of the Shaker cognates in the RACE series was Kv3 (Shaw). All of the techniques utilized for the previous two genes were applied to produce two RT-PCR products and three new UTR exons. One transcript begins with the first coding exon and has a 5' UTR of only two bases, while the other includes three exons

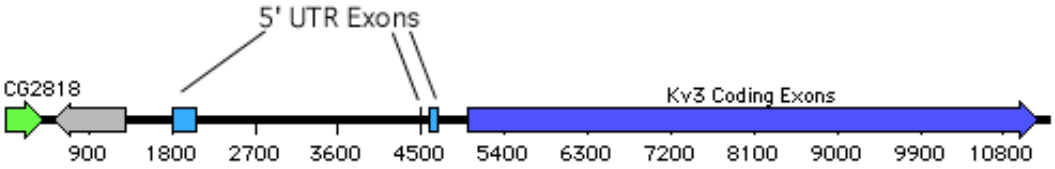


Figure 9: Map of the Genomic Region Around Kv3 (Shaw)

Three additional exons were identified in the 5' UTR of Kv3. These are shown as light blue boxes. The coding region is in dark blue. All three new exons are found in one transcript, while the other RT-PCR product began with the first coding exon. The grey arrow represents an EST match on the opposite strand and the green arrow is the next gene upstream of Kv3 as reported in GenBank. The numbers represent the location on the genomic clone.

Mapping the regulatory region of the Kv3 gene showed it to be a good candidate for reporter gene construction. We sub-cloned 6 Kb upstream of the first coding exon and used this to build the remainder of the reporter constructs. The total size of the construct was just over 16 Kb, which was a normal transformation vector size to inject into embryos. The presence of two transcription start sites might also reveal a tissue or developmental stage specificity for each promoter. In addition to serving as a template for cloning the same region from other species to begin a sequence analysis of conserved regions, it also was used to identify a previously unreported fourth member of the Shaker family.

Kv3.2

In the course of the sequence analysis of the genomic DNA from several species (Chapter 4), anomalous results would arise when aligning coding sequences near the 5' terminus of the Kv3 transcript. In particular, a near perfect match was found in two locations at opposite ends of the chromosome arm 2L. This was later found to be a novel member of the *Shaker* family of ion channel genes (McCormack, 2003). It was reasoned that if this was the product of a gene duplication event which would produce a

closer relative to the Kv3 gene than the other members, then it may be possible to compare the promoter and regulatory sequences between the siblings.

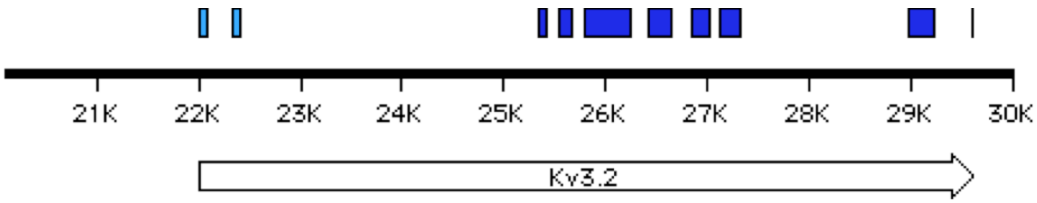


Figure 10: Predicted Exon Structure of CG4450 (Kv3.2)

In 2005, CG4450 was annotated in the Berkeley Drosophila Genome Database as a potassium ion-channel with the exon structure shown above. The dark blue boxes above the sequence (heavy black bar) represent coding exons and the light blues are putative UTR exons based on computer models for splice acceptor and donor sites.

A 5' RLM-RACE was performed and two distinct products were produced from adult mRNA. The first product mapped to the first coding exon in the sequence and only the second predicted UTR exon. The extent of this exon was increased 16 bases 5' to a transcription start site between the two current UTR exons. It also adds a new translation start site in frame with the rest of the coding sequence. The second product included Exon 2 and another exon, that is much further upstream from the first coding exon, and beyond two other annotated genes, CG4438 and CG3752.

The RACE was repeated three times and the same set of products was reliably amplified in a series of temperature-gradient PCR reactions. These results show that there are two transcription start sites— as in each of the other Shaker cognate genes— and that one TSS is 19 Kb upstream of the first coding exon. There are two interstitial genes between the first and second exons of this transcript confounding a comparison of sequence with other members of the gene family.

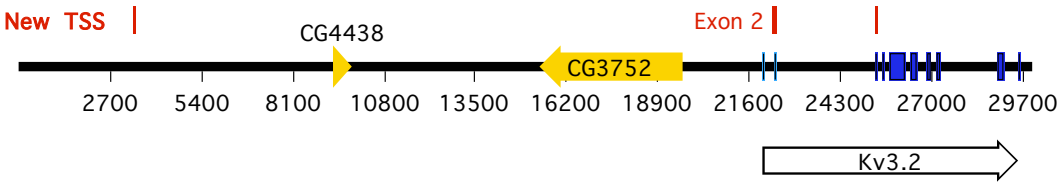


Figure 11: RACE results for Kv3.2

Two products were generated in the 5' RLM-RACE reaction. The smaller included the first coding exon (small red bar to the right) and the next upstream exon labeled Exon 2. This exon is larger than the computer-predicted exon. A second transcript contained a new exon and TSS 19 Kb upstream of the coding region of the gene. The dark blue boxes are coding exons in a predicted open reading frame. The new exons are in red above the sequence bar. The two yellow arrows denote other genes annotated in the region. Both lay between the TSS's for Kv3.2. The 5' most predicted exon was not found in either transcript.

SUMMARY

The promoters found upstream of four members of the Shaker ion channel family were mapped. The transcription start sites were identified along with 5' UTR exons using modified RACE reactions. This showed that Kv2, Kv3, Kv3.2 and Kv4 each have two promoters. It also showed that there is a large size differential between the control regions, with Kv2 over 30 Kb, Kv4 close to 18 Kb, and Kv3 around 6 Kb.

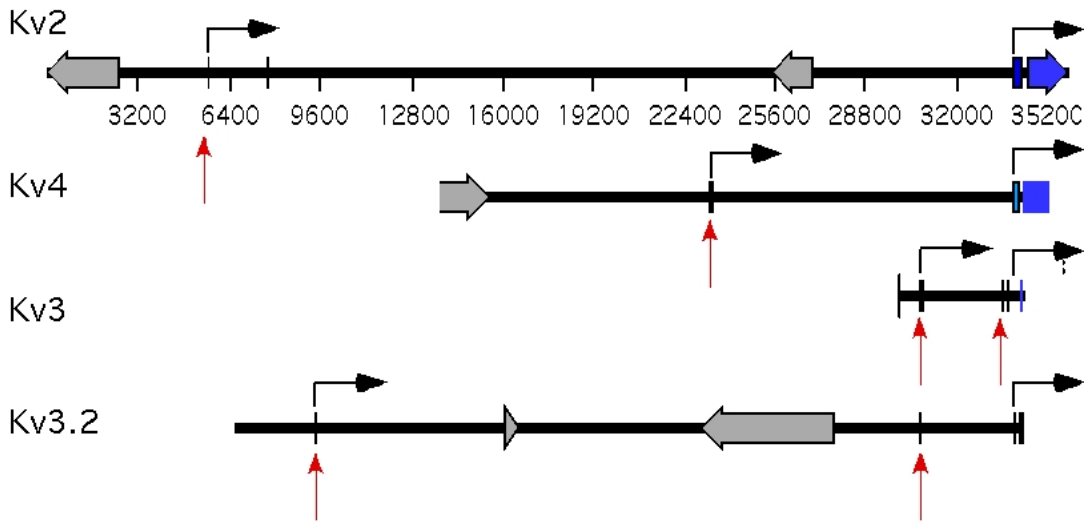


Figure 12: Genomic Map of the relative sizes of the Shaker cognate control regions

Schematic of the control regions shows that the Kv2 control region extends a minimum of 30 Kb, while the Kv4 control region is delineated by the most 5' promoter 14 Kb upstream from the coding exons and the next upstream gene over 18 Kb away. The Kv3 control region proved to be the most tractable, less than 7 Kb. The red vertical arrows point to 5' UTR exons which define the transcription start sites of promoters. The L-shaped arrows above the sequence bars denote transcription start sites. The grey arrows are neighboring genes. Left facing genes are on the opposite strand.

Mapping of the full-length cDNA to the genome reveals several problems that face functional analysis of a control region. Firstly, sheer size can be prohibitive. Cosmid-based transformation vectors, such as CoSper, have been produced for

generating libraries with 10-100 Kb (sometimes larger) inserts, however these have proven to be problematic with extremely low efficiencies as fly transformation vectors (Tamkun et al., 1992). Such large inserts can become cumbersome to generate a well-directed deletion series as they are rife with endonuclease sites in both the insert and vector, not to mention that they would simply require a large number of deletions to screen for critical elements. While certainly a workable endeavor, it was reasoned that the other two members of the *Shaker* cognate group might be more easily cloned and transformed. Another issue is that, as in the case of Kv2 and Kv3.2, other genes along with their commensurate control regions are intertwined in the region concerned. This could confound which of the regulatory elements identified are dedicated to the target gene.

Chapter 4: *in voltro* (sequence analysis)

Given raw sequence for a non-coding region of the genome, the current computational tools available can provide only a rudimentary set of sign posts indicating areas of interest. Sequencing of the same genomic region from multiple species combined with the RACE results provided the basis for a more in-depth *in voltro (in silico)* and *in vivo* analysis. Within the final months of the completion of this work more than 18 insect genomes were nearing completion of draft releases, with many more planned. The utility of whole genome alignments has been profoundly realized and the resources for such projects expanded globally. Whereas annotation and analysis of the regulatory regions of single genomes has been problematic, multiple-genome alignments reveal otherwise hidden relationships and open the door for many avenues of investigation in the future.

GENOMIC CLONING

Genomic clones provide material for both sequence analysis (this section) and the building of reporter constructs (outlined below and discussed in the next section). Once the extent of the control region was identified with the RACE data, the actual genomic material from *Drosophila melanogaster* was needed to build expression constructs for an *in vivo* assay. These clones were also sequenced for treatment with bioinformatics tools. At the same time, the same region from other Drosophila species was cloned, sequenced and compared. In reality, both the cDNA and genomic cloning was done in parallel. Sometimes the sequence from the genomic project was needed to complete the cDNAs and vise versa. Ultimately, all three aspects of the project are interdependent.

Genomic clones and contructs from *D. melanogaster*

Kv4

Genomic clones from the control regions of both Kv4 and Kv3 were produced. In the case of Kv4, a BAC clone existed in a repository at the Berkeley Genome project. This was obtained and a series of PCR reactions using the high-processivity form of Taq polymerase (Elongase, Invitrogen) were performed. The 6-8 Kb fragments were cloned into the transformation vector pPTGAL. These were then injected into flies for in vivo analysis.

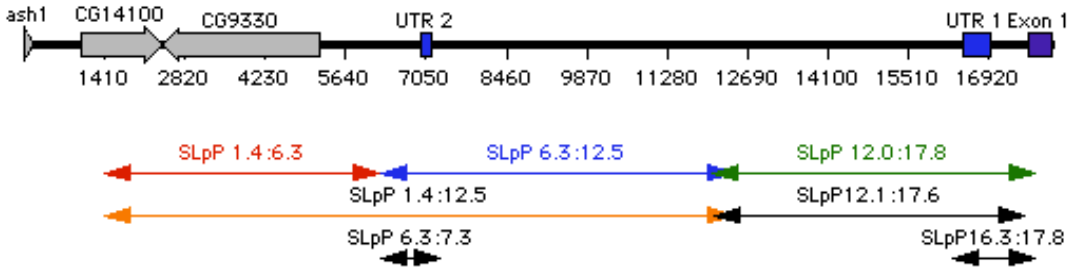


Figure 13: Map of the Genomic Clones in the Kv4 Control Region

This figure shows the cloning scheme for Kv4. The large (~18 Kb) region was divided into thirds and cloned into an expression vector. A total of seven expression constructs were produced. Those in color (red, blue, green and orange) were injected and transformant flies collected and crossed to balancers. The three black constructs were injected and embryos surveyed for transient expression. The dark blue exon is the first coding exon. The two lighter blue exons are in the 5' UTR and mark the two transcription start sites. The grey arrows are matches to ESTs and the well-described *Ash1* gene.

Since no *in vivo* data was produced with the Kv4 construct lines, the clones are discussed here in terms of their sequences and conserved elements, rather than in the next section (*in vivo*). With the Kv4 constructs, the region was divided into thirds. The resulting Red, Blue, Green and Pumpkin fly lines showed minimal expression of the reporter gene against the background fluorescence. This pPTGal vector is an enhancer

trap-type transformation vector utilizing the Gal4-UAS system. Transgenic lines were crossed with both UAS-ßGal and UAS-GFP lines, but there was no distinguishable expression in either case. The pumpkin (orange) line was produced in the hopes that greater coverage would combine necessary elements for expression, but there was still no detectable expression in the nervous system where Kv4 is expected. The "minimal promoter" black constructs were generated with the intent of looking for transient expression in embryos, but none was detected. Later, personal communication with the donor of the vector revealed that others had experienced similar difficulty with known enhancers failing to drive expression of Gal4 from the minimal heat-shock promoter. .

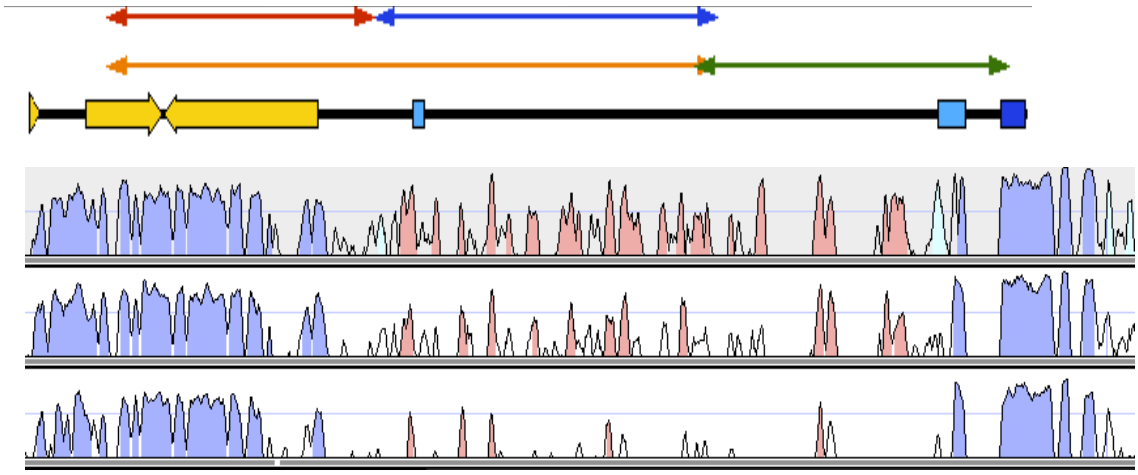


Figure 14: Simplified Kv4 Construct Map with aligned conserved blocks.

The top figure shows the positions of the four primary constructs attempted with the Kv4 control region. The colored double arrows correspond to the fragments described in the previous figure and in the text below. The thick arrows on the sequence bar represent Kv4 (coding exons: dark blue; UTR exons: light blue) and neighboring genes in yellow. The lower figure is an AVID alignment (described in the next section) showing relative conservation of sequence between *D. melanogaster* and *D. erecta*, *D. pseudoobscura* and *D. virilis* species (top to bottom). The purple shaded regions are coding sequence, while the pink shaded areas are more than 70% conserved.

The arbitrary division into thirds may have affected the expression of the reporter gene. It is possible that the regulatory region for Kv4 is not modular. That is to say, that all of the elements (or, at least some elements in each fragment) must be present in order for transcription to occur. It may be that two or more interdependent and requisite elements were not included on each of the "thirds" constructs. An arbitrary guess is that some element in the green construct is critical along with another element in the blue or pumpkin constructs. It may also simply be true that the critical elements for expression of this gene lay outside the region that was cloned. It is interesting to note that there is far less conservation of non-coding sequences in the cloned fragment between melanogaster and *D. erecta* or *D. pseudoobscura* than that found in Kv3. This might indicate that there are regulatory sequences elsewhere, yet to be identified. For these reasons, the regulatory region of Kv3 was selected as a good candidate for a sequence and functional analysis of the regulatory region.

Kv3

The problems faced with the Kv4 gene were absent with the Kv3 clones. The entire region was cloned as a single piece and inserted before a reporter gene. This provided a 'truer' assay of the regulatory elements found in upstream sequences because they interacted with the intact core promoter. The reporter gene is expressed as a direct result of the expression of the initial Kv3 fragment. The Kv3 genomic clone was obtained using PCR primers to the third coding exon and the upstream gene CG2818. The resulting 6.3 Kb PCR fragment was subcloned and then a second round of PCR using primers with an adapter sequence was performed. The resulting fragment could be digested and inserted into a modified transformation vector, pCaSper, such that the reading frame of the Kv3 (Shaw) coding fragment was in-frame with that of the lacZ

reporter gene forming a fusion protein. One possible problem with this series of constructs was if regulatory elements existed outside of the cloned region. This was tested by comparing what was considered a full-length reporter construct (the purple construct in the map below) to an *in situ* analysis of the expression pattern of Kv3 (see *in vivo* data). The constructs are given color names to simplify identification. The other name given is a nomenclature to identify the extent of the construct. For example: SWpC defines this as a Kv3 Shaw fragment in the pFriendlyCaSper vector. The numbers 1.1:5.6 correspond to the fragment from the first genomic clone. SWpC1.1 (4.1∂4.9)5.6 is a genomic fragment with a deletion from the map coordinates 4.1 Kb to 4.9 Kb. At the time that this naming convention was established, four separate vectors and genomic fragments from four genes were in production.

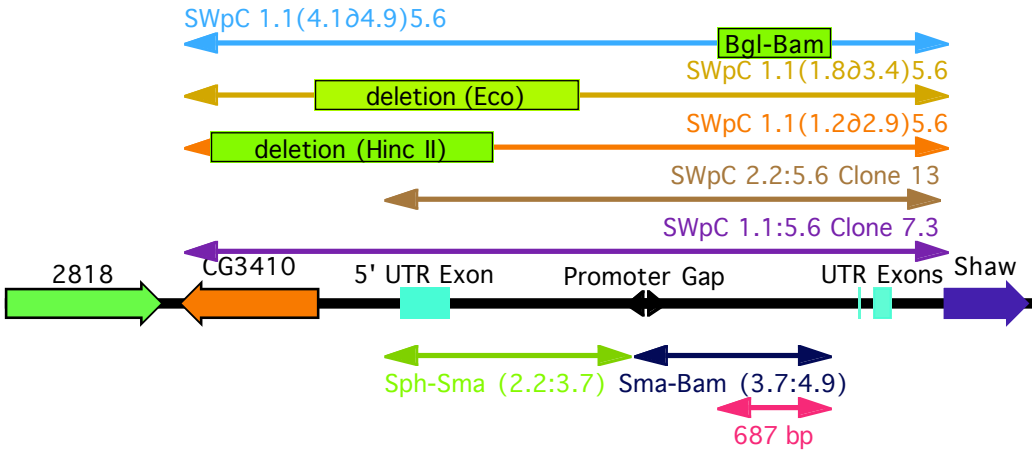


Figure 15: Map of the Genomic Clones in the K3 Control Region

This figure shows the cloning scheme for Kv3. The entire 6.3 Kb region from the third exon of the CG2818 gene (green arrow) to the third exon of the Kv3 (*Shaw*) gene was sub-cloned into pCR-TOPO2.1. A total of 12 expression constructs were produced and eight used to transform flies. Those above the base map are in the pFriendlyCaSper modified reporter vector; those below are in the pHPelican enhancer-trap vector. The green boxes in the upper constructs represent deleted regions. The black diamond marked Promoter Gap is a region of low conservation between species. The dark blue exon (labeled *Shaw*) is the first coding exon. The lighter three blue exons are in the 5' UTR. The two transcription start sites are at the upstream exon (labeled 5' UTR Exon) and dark blue exon (labeled *Shaw*). The orange block arrow (CG3410) is an annotated gene in Release 3 of the *D. melanogaster* genome found on the opposite strand.

Genomic clones from other species of Drosophila

Primers to coding exons were designed and used to amplify a genomic fragment from 6 different species of Drosophila: *D. sechellia*, *D. pseudoobscura*, *D. mulleri*, *D. immigrans*, *D. hydeii*, and *D. virilis*. These were cloned into a standard TOPO vector and sequenced. Primers specific to each species were then designed and used as the gene-specific downstream primer in RAGE, vectorette, and degenerate-primer genomic cloning reactions.

Kv4

The details of the cloning of the genomic fragments are described in Materials and Methods and the *in silico* portion of the Results. In brief, the use of degenerate PCR primers to clone the control region in 4-6 Kb fragments from Kv4 proved fruitless, as there was too little conservation of sequence between species. Primers were designed to the *Ash1* gene and long range PCR was optimized to clone the entire 18 Kb region from four species. Two more clones were obtained by screening blotted libraries from *D. erecta* and *D. virilis*. These were sequenced and aligned using clustalW and blat-slagan programs. The results are shown in the Figure 14.

Kv3

The cloning of the genomic region above Kv3 was problematic. Primers designed to the CG3410 gene produced no results in any species but *melanogaster*. Eventually, it was discovered that this gene is not conserved in the other species of fruit flies (later, it was found in *simulans,* but this species is so closely related to *D. melanogaster* that its transcriptional control region is essentially identical.) The fact that CG4310 is not found in the other *Drosophilid* species required the use of other techniques for obtaining

genomic clones. The length of sequence to the next upstream gene was too great for standard Taq polymerase to amplify. Additionally, the 3' region of this upstream gene is mostly UTR, and therefore highly variable, making the design of degenerate primers difficult. Both RAGE and Vectorette yielded clones from 6 other species.

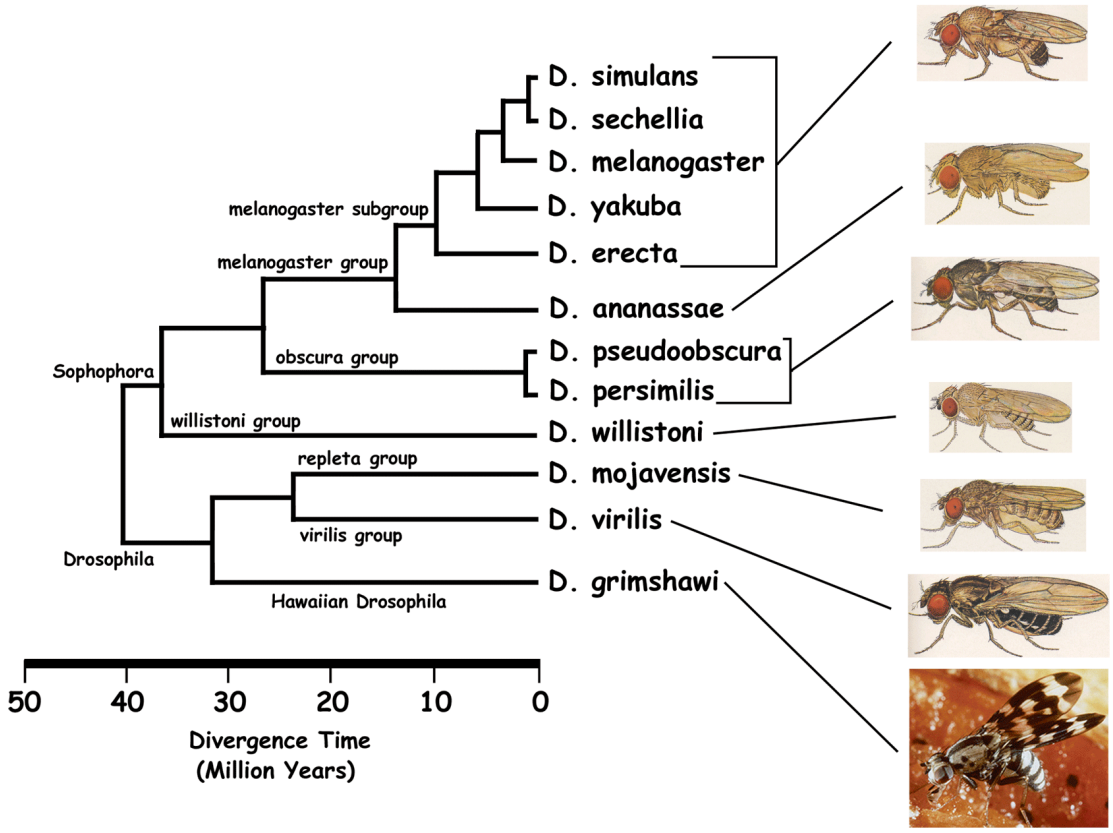


Figure 16: The phylogenetic relationships of the sequenced genomes

The Drosophila genus represents a varied collection of nearly 1000 species. *Melanogaster* has been a model organism for more than a century, but with the recent sequencing of its genome and new comparative tools available, other species have gained attention. The genomic clones for this study span from close relatives within the *melanogaster* group to more distant cousins in the *virilis* group.

Summary

The genomic sequencing project yielded the regulatory sequences for Kv3 and Kv4 from a total of 6 species of *Drosophilids* in addition to *D. melanogaster*. Of these, the sequences from *D. virilis* appeared to have sufficient conservation in the control region to highlight specific blocks, while the sequences from *D. simulans* and *D. sechellia* were far too similar to distinguish blocks of conservation. Other sequences obtained from erecta, mulleri and pseudoobscura appear to have an intermediate degree of conservation in concordance with their phylogenetic relationship. Subsequent genome project releases have confirmed the sequences from *D. pseudoobscura* and *D. virilis*. *D. immigrans* and *D. hydeii* have yet to have their genomes sequenced, but the degree of conservation apparent with a first-pass alignment places them beyond *D. virilis* in terms of evolutionary distance.

The need for chromosome walking came to an end with whole-genome shotgun cloning into high-capacity vectors and large-scale sequencing. During the course of this project, the *D. melanogaster* genome was released and corrected over the next four years. Several other genome sequencing projects are in force and the data readily available on the Internet.

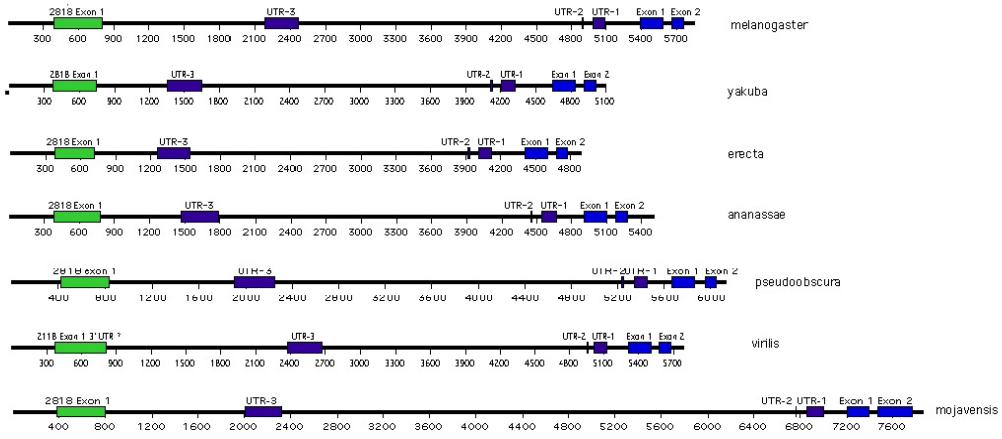


Figure 17: Comparison of Kv3 control region from seven species

Genomic DNA was cloned and sequenced from other Drosophilds for comparison of the general structure of the gene, synteny, and exon arrangement. All show the same three 5' UTR exons and the next upstream gene (CG2818). The coding regions are conserved to over 98% at the nucleotide level. The UTR's are conserved $\leq$ 74%. There are regions between these non-coding exons that are over 85% conserved. These are the sequence fragments targeted in the cloning scheme (see Chapters 4 and 5)

Once the sequences of the similar regions were obtained, they were aligned using Slagan or AVID and conserved blocks annotated. These blocks and regions of very little conservation were used to guide the design of expression constructs. They also served as a basis for the core promoter analysis and the transcription factor binding site analysis. There is an over-all conservation of exon structure conserved throughout all the species studied with some variability in the distances between the UTR exons. Sequence analysis reveals that while exon boundaries are conserved, the actual sequence of the UTR exons was not.

GENOME STRUCTURE AND SYNTENY

The Kv3 gene and the next upstream neighbor, CG2818, were both conserved in the genomes of all the *Drosophilid* species used. However CG2818 was not found to be in the same chromosomal location in the species from other orders such as the silk worm and honey bee. None of the 5' UTR exons or any other sequences in the regulatory region were conserved in other orders. Within the Dipterans, the UTR exons and control regions are conserved among the Drosophilids, but not the Anopheles mosquito. This established the sequences of *D. virilis* or *D. mojavensis* as boundary organisms for the study of conserved non-coding elements (data shown below and other following figures).

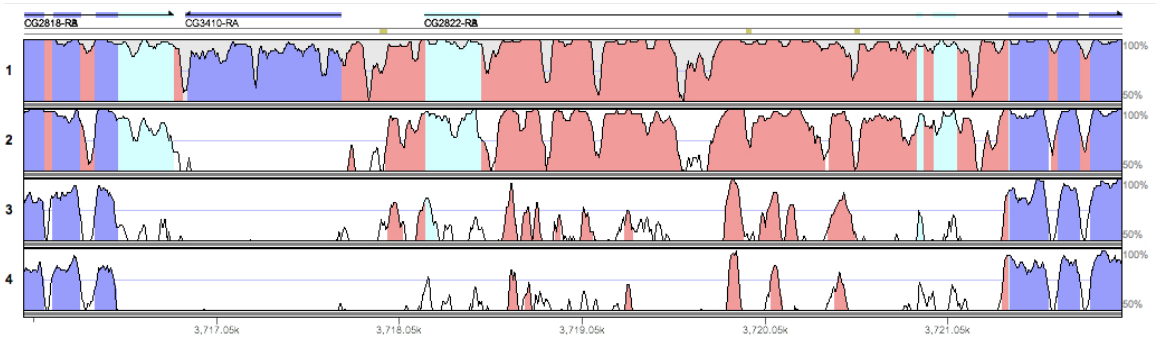


Figure 18: A four-species genomic alignment of the Kv3 control region.

This is an AVID-Vista plot (described below) of the span of genomic sequence from Kv3 to the next conserved gene upstream on the same strand. The peaks in the plots represent the relative degree of sequence conservation between *D. melanogaster* and (1) *D. simulans*; (2) *D. yakuba*; (3) *D. pseudoobscura*; and (4) *D. mojavensis*. There is a general loss of conservation from plot 1 to 4. Blue denotes coding sequences on the base strand (*D. melanogaster*) and pink is any region of greater than 70% conservation. In particular, the CG4310 gene is not conserved at all. It was later identified in the closest *D. melanogaster* neighbor: *D. simulans*, but in no other species. This is a recent insertion of a gene or pseudo gene into this intergenic region.

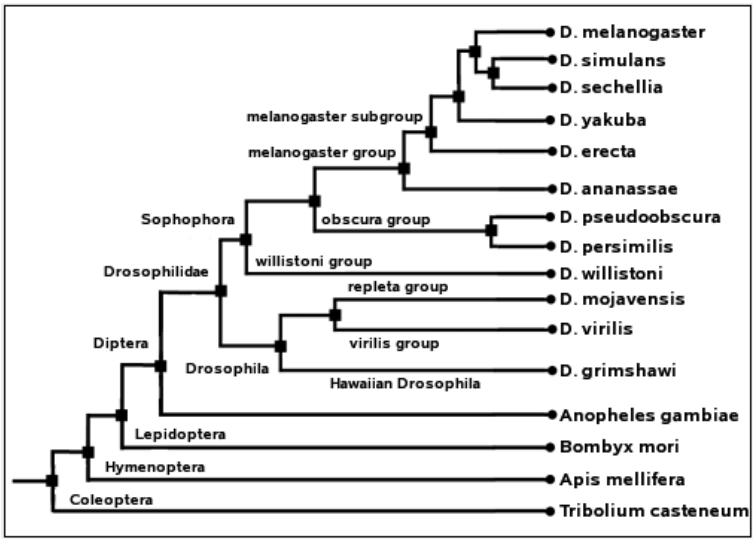


Figure 19: Cladogram of some Insecta class members

> The assortment of genomes that have dedicated sequencing projects has expanded from 2-3 in the late 1990's to dozens today. Several have serious clinical applications such as the *Anopheles* mosquito in malarial studies to the economically important *Bombxy* silkworm moth. Many other species affecting agriculture are coming into the spotlight also. In genetics studies, the distant members in the *Dipteran* order along with other orders have little conservation of synteny or regulatory sequences, but show cases of gene duplication over scales greater than 40 million years.

The concept of evolutionary divergence or distance between species is evolving rapidly itself. In the past, almost all comparisons were performed in coding sequences. However with the advent of whole genome sequencing, comparisons of non-coding sequences has shown that some of these regions are more conserved than coding exons. This will alter some of the time lines and the 40 million year figure for divergence time in the Drosophila may change. Another issue is that selective pressures on insects cause rates of variation substantially different from that of mammals such that conservation studies between taxa cannot be readily translated across families. We are only concerned with the elements that can be identified as conserved within the Drosophilids.

ALIGNMENTS

clustalW

Clustal is a global alignment algorithm that takes either nucleotide or amino acid sequences as its input. It can align multiple sequences and scores using simple log-odds values in a sliding window. It incorporates gap penalties, with the assumption that the sequences to be aligned are particularly similar in size and identity. This is is a useful algorithm for coding regions, but is problematic with large stretches of genomic with non-coding sequences. The blat anchoring and slagan long range alignment algorithms of AVID are better for those types of sequence.

A blast with a coding fragment from Kv3 (transmembrane segments S4-S6) produced multiple hits in the *Anopheles gambiae* genome and the *D. melanogaster* Release 3. The hits to Kv1, Kv2 and Kv4 were expected in the fruit fly, however a hit to the gene annotated CG4450 scored higher than to either Kv2 or Kv4. This then could be a new member of the *Shaker* cognate family. We designated this as Kv5, but in a subsequent study by McCormack *et al*, they designated the hit to the duplicate Kv3 as Kv3.2 (see figure below). We then searched the genomes of the silkworm and honey bee databases and found only the single hit to Kv3. This suggests a gene duplication event in the Dipterans, not found in the other insect genomes. However, these genomes are not considered complete and further curating of sequences may reveal an earlier duplication. Even so, this means that there is another potassium channel in Drosophila that is more closely related to Kv3 than other members of the cognate group. This may prove to be useful in the future for analysis of common elements in their respective regulatory regions.

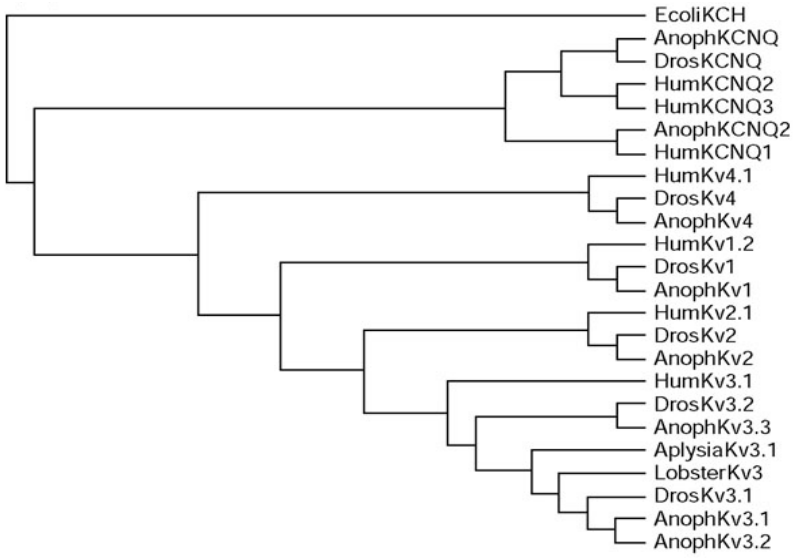


Figure 20: Phylogenetic tree of voltage-gated channels.

This is a partial tree of the voltage gated channels found in several species. The Kv3 genes are listed at the bottom showing three Kv3 channels in Anopheles and two in Drosophila. This was confirmed with searches through other Drosophila genomes extant and a second mosquito genome: *Aedes aegypti*.

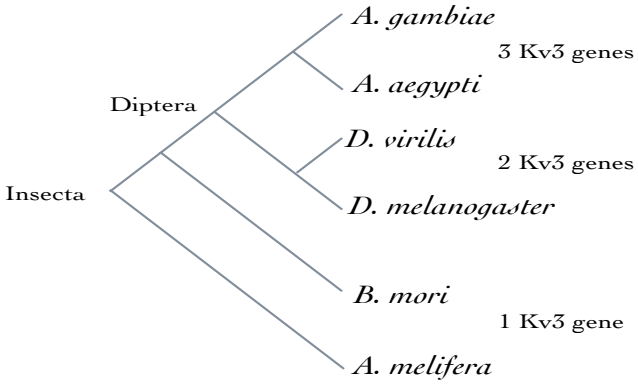


Figure 21: Kv3 duplication events

Three Kv3 genes are found in the mosquitos (*A. gambiae & A. aegypti)*. Two in the fruit fly and one in other insects currently. This may be a case of gene duplications specific to the Dipterans. Why this would be specific to the Kv3 and not other members of the Shaker cognates is not known, but it does raise the possibility of a comparative analysis of the regulatory regions of closely related genes.

AVID-Vista alignments

AVID is an alignment tool capable of comparing whole genomic sequences. It employs a strategy that begins with local alignments of highly conserved sequences to serve as 'anchors'. This first pass uses the BLAT algorithm and variants (Kent, 2002). It then applies a global alignment computational algorithm (SLAGAN) to find conserved blocks between these anchor sequences. This tool has proven particularly useful in finding orthologous genes in multiple species (Bray et al., 2003). To complete the tool set, Vista is a viewer to display the results of the alignment. Originally a freely available binary and Java applet, Vista has grown into a suite of database and genome search tools that allow for primary search and annotation of the genome sequences (Couronne et al., 2003).

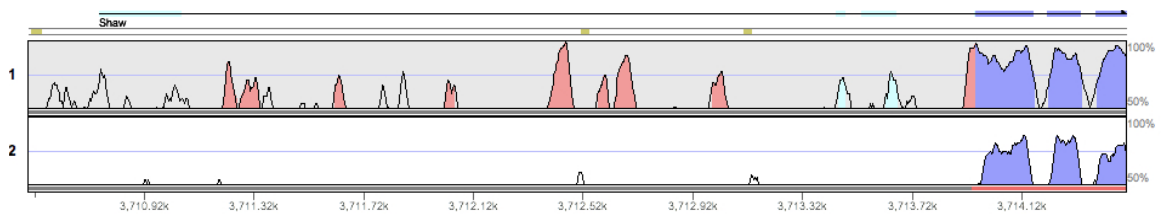


Figure 22: Comparison of promoter control region with *Apis mellifera*

This is a locally-run slagan alignment with the honeybee genome assembly (January 2004). The panel labeled (1) is the *D. virilis* alignment from the previous six-species run, used as a reference for conserved non-coding sequences. The second panel (2) is the orthologous sequence from *Apis mellifera*. This illustrates that while the coding sequences (shaded purple) are still conserved in the honey bee, the upstream regions are not, including the UTR exons.

The coding region of Kv3 was aligned to the genomes of several insects: the Dipterans: Drosophila fruit flies and *Anopheles* mosquitos; and representatives from two closely related orders, Lepidoptera (silk worms: *Bombyx mori*) and Hymenoptera (honey bees: *Apis mellifera*). Although the coding sequences were well conserved— greater

than 70% across Orders, and over 85% within the *Dipterans*— the upstream sequences, including promoters and UTR exons were not conserved enough to annotate any blocks. This served as a logical limit to the range of species to compare in the search for regulatory elements. It is important to note that once sufficient information is known about the regulatory sequences of other Orders, elements found to have a common function may serve as "rosetta stones" for identifying transcriptional themes.

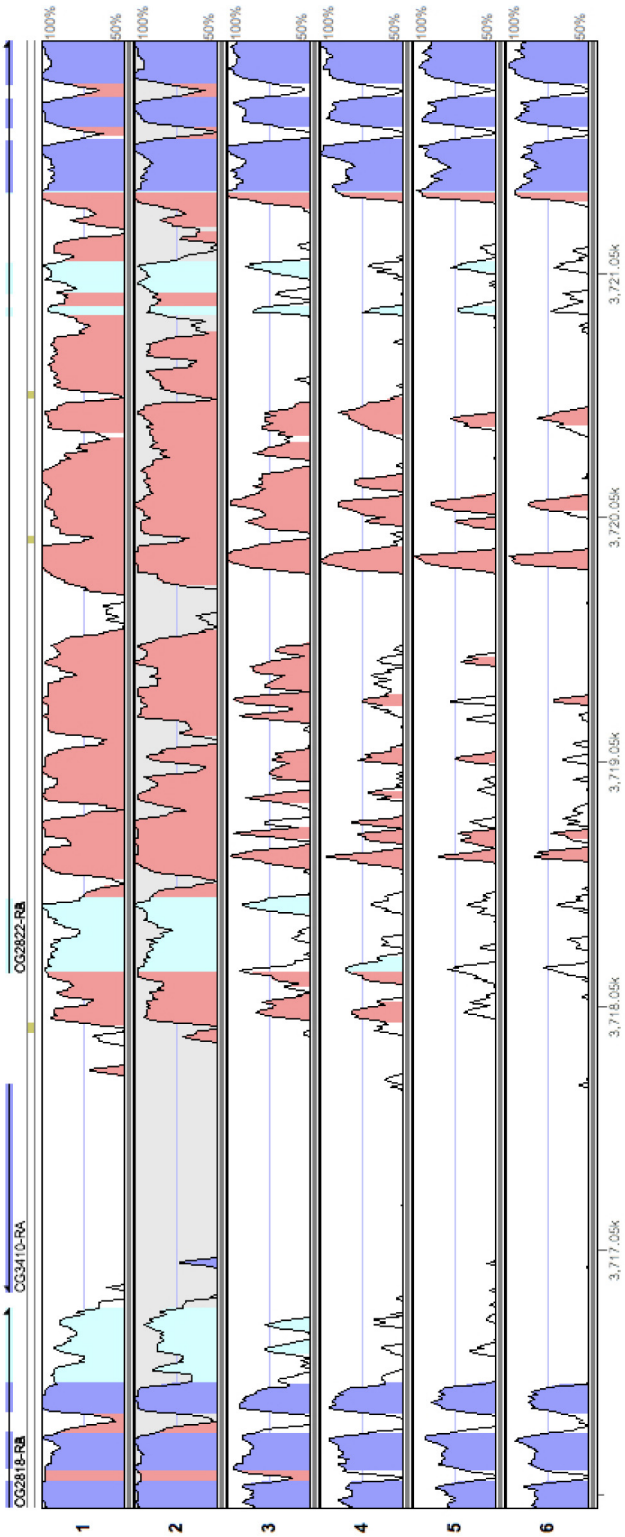
By setting an arbitrary cut-off value of 50% for conservation of the non-coding expressed sequences, the *D. virilis* sequence became a "boundary" set of data to use in comparisons. At this evolutionary distance, UTR exon splice donor and acceptor sites are still sufficiently conserved to map reliably. The conservation degrades towards the center of the exons. Core promoter sequences are not conserved, however highly conserved blocks of sequence that are equidistant both from each other and a neighboring transcription start site are found. This implies that the functional constraint on transcriptional regulation lay not at the core promoters, nor in the sequences of the UTR exons, but rather in the intermediate regions. These are likely transcription factor binding sites or motifs that are recognized by chromatin remodeling proteins either for direct modification of nucleotides or as benchmarks that call for the modification of bound histone molecules.

Figure 23: Vista plot of an AVID alignment of 6 species of Drosophila

The annotated baseline sequence (that which all other species are aligned to) is the line at the top of the figure is *D. melanogaster*. The six blocks below are the alignments to (1) *D. yakuba*; (2) *D. erecta*; (3) *D. ananassae*; (4) *D. pseudoobscura*; (5) *D. virilis*; (6) *D. mojavensis*. The graph plots the degree of conservation between species in a sliding 50 bp window. Purple represents regions of coding sequence; light blue marks UTR exons, while the pink shaded areas denote ≥70% conservation of sequence over 100 bp. The percent conservation scale is printed on the right of the graph with 50% as a minimum cut off.

The baseline sequence is from Chromosome 2L (3,716,000-3,722,000 bp) of the April 2004 release of the *D. melanogaster* genome. The 3' half of the gene: CG2818, the entire EST match to CG3410, and the control region and first three coding exons of Kv3 (CG2822, Shaw) are included.

There is an obvious region of low conservation between the Kv3 UTR exons around the 3,719,750 bp mark. This is denoted as the "Promoter Gap" on cloning maps. Between this gap and the first coding exons are three distinct peaks where the conservation of sequence exceeds 70%. These are the sequences targeted in the cloning scheme in the *in vivo* section of the report.


CG2818-RA
CG3410-RA
CG2822-RA
100%
50%
3,717.05k
3,718.05k
3,719.05k
3,720.05k
3,721.05k
1
2
3
4
5
6

Phylogenetic Footprinting

There is a rapidly-growing set of techniques designed for identifying conserved non-coding sequences (CNS's). Recent mammal studies have shown that some regions of the genomes are far more conserved than coding regions or UTR exons. These studies were conducted to address the differences in the organisms in cases where there is insufficient variation in the coding sequences to explain the relative differences between the organisms, such as with mice and other mammalian model animals (Sironi et al., 2005; Wasserman et al., 2000). Regions of "ultra-conservation" that suggestion strong functional constraint were identified and it was suggested that their regulatory role forced a greater influence of negative selection against variation in these sequences than in neighboring coding regions.

As outlined in the description of the AVID-Vista plot, the sequences from assorted Drosophilids reveal conservation of putative regulatory elements that are more conserved than untranslated exons, but less conserved than the coding regions. This allows for arrangement of the phylogenetic tree in closely related groups such as *D. simulans*, *D. sechelia* and *D. melanogaster*. This will become more important when comparing the functional role of specific regulatory sequences. There was an attempt to exchange these sequences in transformation constructs between two species of Drosophila but they failed to drive expression of the exogenous promoter. This suggests that while there is sufficient conservation to recognize these regions in different species, there are other differences in the transcriptional regulation machinery that prevent direct testing for common function of these elements in foreign species (Betran e Long, 2002; Gonzalez et al., 2000).

PATTERN ANALYSIS

Matching to known matrices

Core Promoter Elements

Once the 5' RACE was completed, the exons that were discovered were mapped to the genomic. The intent is to locate the true transcription start sites (TSS). It is here that the sum of all the effects of the regulatory sequences are focused and where the initiation complex must form. In order to build constructs that reflect the nature of endogenous expression, the TSS must be known. However, even persistent efforts at 5' RACE have failed to identify promoters. The foibles of RNA such as secondary structure or rapid degradation can foil a RACE.

In a recent study highlighted in the introduction, Ohler *et al* scanned all the putative TSSs annotated on the Drosophila chromosome arm 2R. A 150 base pair region around each site was entered into the MEME motif-finding algorithm and 10 different patterns of elements were found to be common amongst over 70% of the sites (over 2000 genes) (Ohler et al., 2000). These included the two canonical TATA and Inr-DPE (Initiator-Downstream Promoter Element) type promoters, along with eight other motif clusters. They describe a new type of DPE that is similar to, but distinct from the known form. Certain motifs are more often found in a positional relationship with one another. The common DPE typically is 3' of an Inr site. TATA-less promoters typically have an Inr site. These spatial relationships can be added as a heuristic to MEME as a prior for Position-Weight Matrices in the absence of a background filter (Bailey e Elkan, 1995). These values were obtained from the Ohler study.

To assay the two TSSs found in the Kv3 region, Meta-MEME seeded with the 10 Drosophila promoter consensus sequences was run locally against both the coding and non-coding portions of the Kv3 region. The downstream promoter was evaluated as a TATA (only, no other elements) type promoter, while the upstream promoter at the 5' UTR had three elements found in some TATA-less Drosophila promoters.

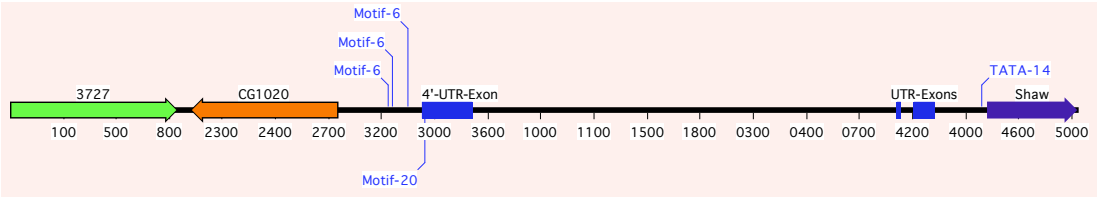


Figure 24: Meta-MEME matches to TSS of Kv3

The TSS found in the first coding exon shows a match to a canonical TATA-box site in a position-correct state. No other elements common to TATA promoters were found. The 5' TSS matched an Inr-DPE (type 10) motif found in a previous study of Drosophila TSS's. Additionally, more than 65% of those new Type-10 DPEs were found associated with a Type 7 motif. Three Motif 7 sites are found upstream of the 5' Inr-DPE promoter.

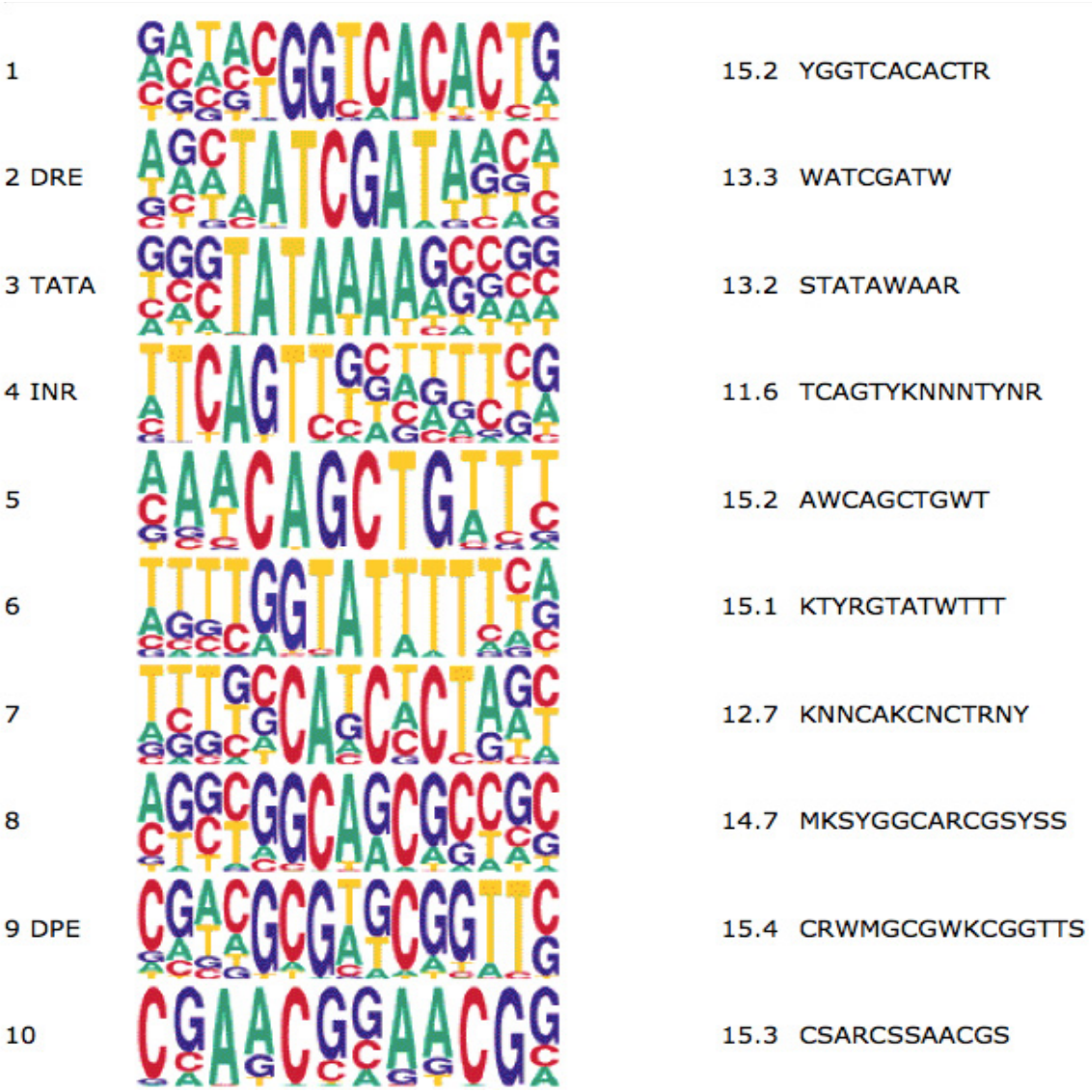


Figure 25: Core promoter elements identified in Drosophila genes.

A MEME run trained against coding and non-coding sequences identified 10 distinct elements localized to the transcription start sites for 2000 genes in one arm of Chromosome 2. The different elements are numbered to the left with known elements listed. The pictograms in the middle column shows the frequencies of each nucleotide where the height of the letter corresponds to the rate of occurrence. Consensus sequences are to the right (Ohler et al., 2002).

Match and Patch and Pseudosequence

The Transfac database was obtained and used to search the regulatory region of all the *Shaker* cognates. The reasoning is that short regulatory sequences that do escape the parameters of alignment programs might be recognized in common by transcription factors. That is, an enhancer sequence that binds factor A in the Kv3 control region might also be found to be bound by A in the control region of Kv3.2— however these two enhancers could have sufficient divergence of sequence to go undetected by current alignment algorithms. The fact that these sequences are short, often less than 10 bases, and that transcription factors will bind to multiple sequences, is sufficient cause to believe that regulatory sequences could go unnoticed with current tools. We sought to search sequence sets in two ways: first search the same region in multiple species, and then search the control regions of functionally-related genes. The database is heavily biased towards vertebrate transcription factor sites with only 38 insect matrices (some from *D. pseudoobscura*) included. We expanded upon this set by hand-curating an additional 62 matrices from clustalW alignments of data from the literature. In the case where only a single sequence string was available, this was entered as site in the Patch database.

Match (matrix based search) and Patch runs were generated and the results converted to single letter codes in strings that are spatially correct (null characters were used as intervening bases to establish correct distances between single letter representations of the Match/Patch results). This was done with an in-house perlscript. This "pseudosequence" was then concatenated into a multi-sequence fastA file which was submitted to a modified clustalW called tfclust. This generates alignments of pseudosequence in asn format which can then be viewed with graphical sequence

analysis programs. The intent is to find patterns or clusters of sites that are in common between related genes. The alternative is to find similar patterns or clusters of sites in non-conserved sequences of homologous sequence between species. After several runs with broad spectra of constraint parameters no obvious results were found. There is still no reliable automated method of determining which Match or Patch "hit" is significant. This requires a human involvement that exceeds the capacity of the resources for this project, but it is clearly an idea that is fomenting in the genomics community. The Lawrence Berkeley Lab offers an equivalent form of automated alignment to known binding sites with multi-species SLAGAN alignments, however the largest drawback of that approach is the dearth of established transcription factor binding sites, and for that matter, whether the sites that have been described reliably reflect the actual sequences recognized by regulatory proteins. Given that the insect transcription factor data base has fewer than 100 entries, the vast majority of these from developmental studies, there is likely to be at least an order of magnitude more factors to be added to the database. It has been estimated that 8-10% of the proteins coded in the genome are specifically for regulating the transcription of other proteins, yielding more than 1000 in Drosophila alone (Tupler et al., 2001; Adams et al., 2000). Perhaps as the quantity of known transcription factor binding sites increases and the quality of the individual matrices improves— that is, the amount of information content reaches a sufficient minima— then such analyses will be possible. Several personal communications with other investigators doing research along these lines have found them to be in agreement that the application is just over the horizon, but not quite available yet.

Searching by information content

MEME

MEME is more fully described in Methods, but in brief, MEME is an unsupervised learning algorithm for finding motifs in a sequence. It uses a Bayesian probability to incorporate prior knowledge into its searches. In a reiterative approach, the localization of motifs can affect the efficacy of finding other motifs within a given sequence. The most recent release adds the capability of considering a background state, which is an assessment of the assortment of sequence elements found in training sets that are either to be excluded (demoted in value) from searches or more heavily weighted. An example would be a set of confirmed coding exons (replete with codon bias) as a method to minimize false positives. Non-coding sequences that are pre-masked for repeats are used as false negative training sets. The output is a set of profiles of motifs in the form of a log-odds matrix. These are then compared to an unknown sequence to score against the matrices to find matches to motifs.

Sequences can be entered singly or as sets. Here a *D. melanogaster* regulatory region sequence was annotated with a motif profile found in a single sequence (against the non-coding background file) that aligns to the conserved sequence blocks found in the multi-species AVID plots. This implies that the conserved blocks not only are under functional constraint to conserve sequence between species, but also that the motif itself is functionally important in that it is repeated four times in the 6 Kb stretch. The expect values, the likelihood that this motif would randomly appear in this sequence, are particularly low.

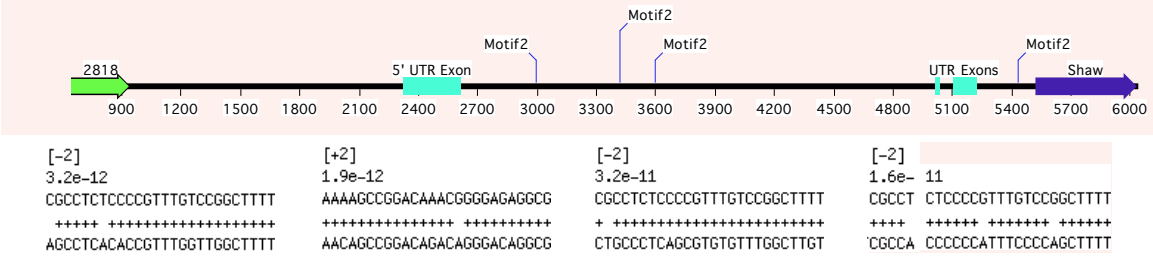


Figure 26: MEME analysis of the Kv3 control region.

Three motif profiles were found in the control region of Kv3. Motifs 1 and 3 had single occurrences with high expect values in the region upstream of the 5' UTR exon. While motif 2 had four hits, each with expect values lower than e^{-10}. The appearance of all four 26 bp hits in 6 Kb with such expect values is likely in less than 1 in 10^{10} iterations of the same sequence in random order.

The triplet of motif sites in the 600 bp run match to conserved sequences "peaks" in the AVID alignment described above. There is little doubt that these are functionally constrained motifs. Determining that function is a much larger task. An alignment to Transfac databases yields no match within reasonably bounds. A larger pool of functional elements or *in vivo* data will be required to understand their role in this region.

McPromoter

Another example of a Hidden Markov Chain search algorithm is McPromoter. It is designed to look for eukaryotic transcription start sites. The system utilizes a background model consisting of "priors", or statistical states for coding and non-coding sequences, and a standardized promoter model that segments the region into six discrete elements. The input sequence is couched against the prior models over a window of 300 bases, which it stepped 10 bases at a time. The respective promoter and background likelihoods are then fed into a neural network, along with likelihoods representing the DNA structure in the six segments (Ohler et al., 2000). A threshold setting can be changed; however as this is lowered, the sensitivity increases inversely with specificity.

A prediction is made for each local maximum above the set threshold for each 10 base pair step of the sliding window. As a statistical system, it does not require that certain patterns must be present, but that the combination of all features is good enough. E.g., even if the TATA box score is very low, there can still be predictions if the other features score well (Ohler et al., 2002). This system was trained in a similar schema to MEME and the Tranfac Match algorithm, with minimization of false positives scored against coding sequences and false negatives against known promoter sets from nearly 100 Drosophila genes.

Given the putative regulatory sequences upstream of the coding region for Kv3 and Kv4, McPromoter predicted a strong likelihood of a promoter 3200 bp upstream from the reported promoter. Subsequent 5' RACE and the *in vivo* studies confirmed this prediction.

SUMMARY

Genomic clones from six species were obtained and sequenced. These were aligned with BLAT-SLAGAN-AVID and plotted using Vista. Regions of high sequence conservation were annotated and used to direct the design of reporter constructs. A gene found in *D. melanogaster* 5' of Kv3 is not found in most other species of flies and is likely a recent incorporation event. The core promoters in the Kv3 control region were mapped and classified using various sequence analysis algorithms to confirm *in vitro* data. A common motif was found in the regulatory region that is highly conserved and extremely non-random in occurrence. Transcription factor binding site recognition trials were attempted, but no meaningful results could be elicited.

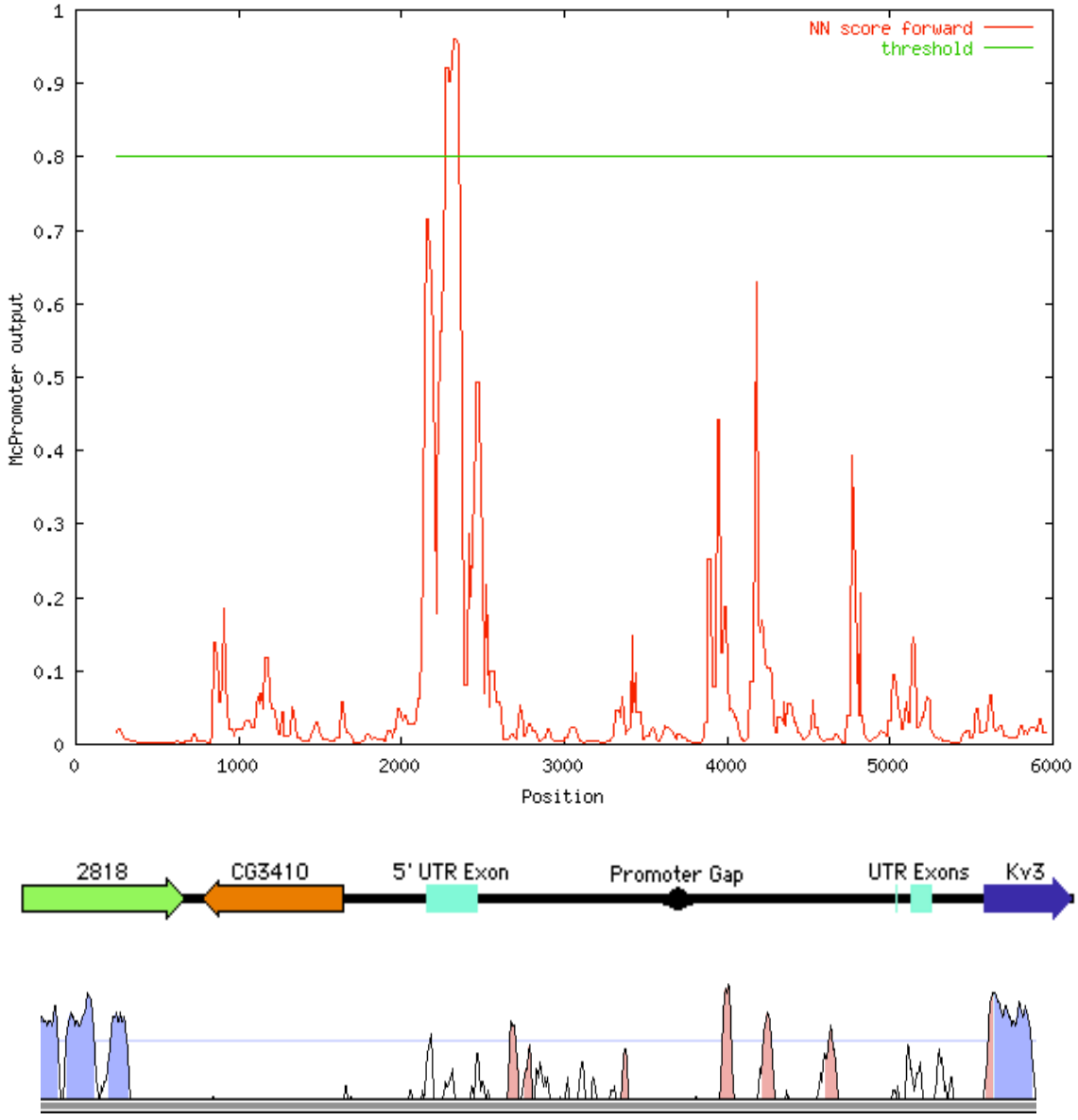


Figure 27: McPromoter analysis of the Kv2 regulatory region.

The trace at the top of the figure is the output from the McPromoter neural network. It screens raw sequence using a HMM chain-rule algorithm trained against known promoters. The super-threshold score locality agrees with *in vitro* and *in vivo* data (annotation of 5' UTR Exon in middle panel). The bottom trace is the AVID pairwise alignment with *D. virilis*, which fails to identify the 5' promoter, but does align conserved blocks with possible regulatory elements around the 4000 bp region.

Chapter 5: *in vivo*

A first step in the functional analysis of a control region is to determine where and when the endogenous gene is expressed. This was done by embryo *in situ* hybridization. In summary, we observed that all three channel genes were expressed in embryonic stage 11 neural tissue. During Drosophila embryonic development, stage 13 (Campos-Ortega) is the time point at which germ band retraction terminates and the CNS and PNS cells differentiate (see figure). The CNS neural tissue comprises a ventral nerve cord (VNC) and the procephalon (brain), while the PNS, is organized into three general organs: the sensory-motor neurons, sensilla, and the stomato-gastric nervous system.

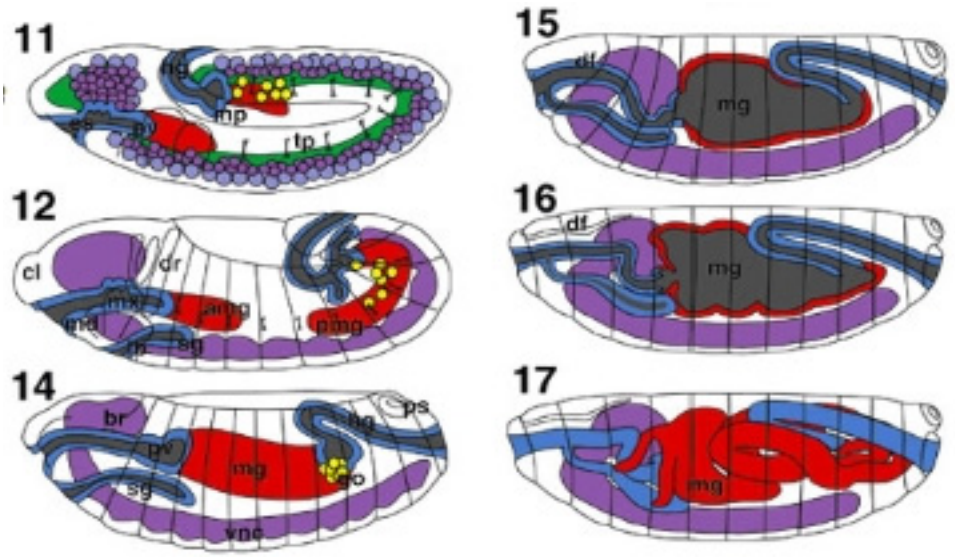


Figure 28: Schematic of embryonic development, stages 11-17.

The developing nervous system of the Drosophila embryo proceeds from stage 11 to 16. By stage 13, the retraction of the germ band has completed and there is a well-differentiated ventral nerve cord (vnc-purple). There are also the supra-oesophogeal ganglia of the brain hemispheres, or procephalon. The stages are easily defined by the progress of the closure of the mid-gut. Most of the images below are from stage 13-15. After: (Holmes e Tjian, 2000)

The expression of Kv3 by stage 13 is primarily in the VNC and procephalon.

There is further expression in the PNS in stage 15-17 embryos (just prior to hatching). After stage 16, deposition of the cuticle makes staining difficult. To show this expression, late-stage embryos were stained with antibodies against ß-galactosidase. Kv3 is also found in the ring gland which is believed to be a neuro-secretory gland composed of the corpora cardiaca, thoracic glands and the corpus allatum. This appears very late in embryonic development in stage 17. There is also some transient staining of the midgut, which is difficult to distinguish from endogenous ß-galactosidase activity.

The endogenous pattern was then compared to the pattern of ß-galactosidase staining in the full-length construct transformants. This construct contains all of the genomic material from the first coding exon upstream, to the next gene 5' on the same strand. The SWpC1.1:5.6 (purple) construct faithfully repeated the pattern of expression as illustrated by the *in situ*. The expression pattern directed by the two halves of the regulatory region are then analyzed with deletion and enhancer trap constructs.

IN SITU DATA

A probe was designed to a gene-specific region of the cDNA. The extra-cellular loops between the transmembrane segments are highly variable and minimized cross-reaction to other channel gene products. The probe for each gene was located in the first extra-cellular loop. An optimal negative control would be a null mutation of the target gene. Preferably one that truncates the message or prevents expression at all, this would ensure that any hybridization seen is genuinely for the target gene and not the result of some cross-reactivity. Since no homozygous deficiency mutants exist for Kv3 or Kv4, a negative control in a null background was not possible. However, sense probes were used as controls to account for non-specific binding of nucleic acids to tissue and these

showed no non-specific binding. The anti-sense probe showed that the earliest expression of the Kv4 and Kv3 genes is in the stage 13 embryo. All expression is focused in the neural tissues of the embryonic brain and ventral nerve cord. By later stages until hatching, faint expression of Kv3 is seen in the PNS and the antenno-maxillary complex.

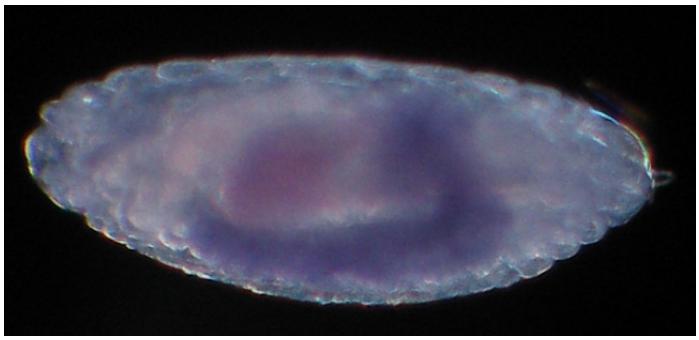


Figure 29: Kv4 *in situ*.

This stage 13 embryo shows distinct staining of the neural tissues in a whole mount preparation. Anterior is to the right and dorsal at the top. Heavier staining is found towards the interior of the procephaic lobe and the ventral nerve cord.

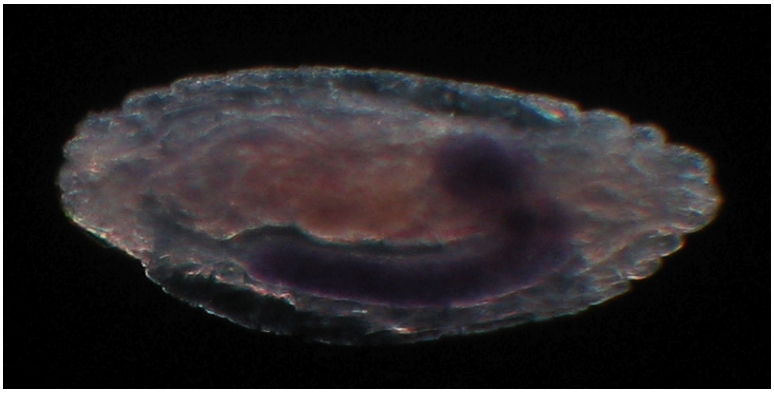


Figure 30: Kv3 *in situ*.

Stage 13 embryo showing the Kv3 expression pattern in the primordial neural tissues. Anterior is to the right and dorsal is at the top. Staining is of the central nervous system and the ventral nerve cord at the bottom of the figure. The pattern is similar to Kv4. It is also similar to that of the full-length reporter construct.

Hodge et al, recently described the expression pattern of Kv3 with *in situ* of late

stage 16 embryos. They report a similar CNS pattern along with distinct staining in the PNS and the antennomaxillary complex (Hodge et al., 2005).

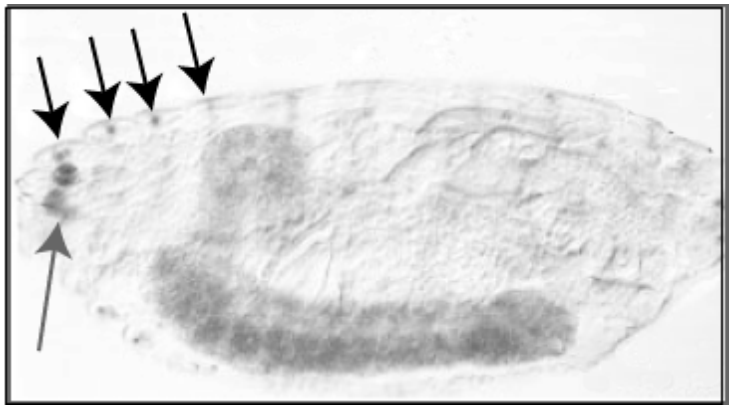


Figure 31: Stage 16 expression pattern of Kv3

in situ of late stage 16 embryo. Expression of Kv3 becomes apparent in the PNS (black arrows) and in the antennomaxillary complex (grey arrow), as well as the CNS. Ventral nerve cord is at the bottom. Anterior is to the left. There is also some faint expression in the dorsal caudal sensilla .

Figure 32: Kv3.2 *in situ*.

Stage 12-13 embryo showing the Kv3.2 expression pattern in the procephalic tissues and the ventral nerve cord. Dorsal is down and anterior is to the right. This pattern is more concentrated in the central nervous system cells and less dense than that of Kv3. There is also some possible expression in the mid-gut.

In addition to the Kv3 and Kv4, an *in situ* of the novel gene, Kv3.2, showed a

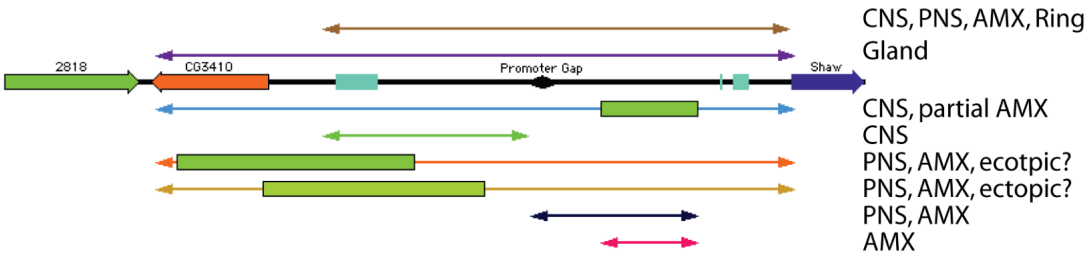
similar pattern in the embryonic stages. This gave hope to the notion that a comparison of the control regions of these genes would yield common elements or arrangements involved in this representative expression pattern in the embryo.

TRANSFORMANT LINES

Once the embryonic expression pattern of the endogenous gene was known, a full-length reporter transgene was built for functional analysis. If this reporter faithfully expressed in a pattern similar to the endogenous gene, there would be some assurance that most, if not all of the cis-regulatory elements responsible for spatial patterning were present. As illustrated in Chapter 3, a series of transgene constructs were then generated using fragments of the control region in an enhancer trap vector to drive expression of the lacZ gene, or the entire control region driving expression of a fusion protein. The latter was then subjected to a series of deletions to derive which portions were critical to expression in either a spatial or temporally specific manner. The decision as to where these deletions were made was driven by data from the sequence analysis of multiple species alignments. Areas that were conserved, alignments to transcription factor binding sites, motifs discovered by MEME, and McPromoter data were used to map a strategy of constructs to isolate certain regions. These were either deleted from the full length vector or inserted into an enhancer trap.

One consideration in building expression constructs is that a small deletion might remove an element that is critical to all expression, whether tissue or stage specific. Additionally, repressor elements might be deleted, producing a sudden increase in ectopic expression. With this in mind, over-lapping constructs were built with the intent of looking for elements that are either repressors or enhancers. Two different vectors were used to specifically address a case where a small fragment might not drive expression in an enhancer trap, but could reveal its function after being deleted in a fusion-gene construct. Another factor that might confound separating discrete regulatory elements is

that a deletion will bring two components into proximity that are normally separated by a stretch of sequence. A change in expression could be misinterpreted if a normally active enhancer is brought under the control of a repressor or silencer. Several of the deletion constructs failed to produce any expression; whether this was due to the case above, or a situation of a missing critical element cannot be determined. Lastly, with enhancer trap vectors there is some leakage, or basal expression from the heat shock promoter (HSP70) that cannot be accounted for with vector-only transformants.



CNS- central nervous system; PNS- peripheral nervous system; AMX- antennomaxillary complex

Figure 33: Summary of reporter construct lines and expression patterns.

At the center of the figure is a schematic of the control region for Kv3. Above and below are double arrows representing the regions cloned into each construct. The green boxes are deleted sections. To the right is a summary of where each construct expresses.

In general, the control region is divided into two sections: from the first coding exon to the promoter gap; and from the promoter gap, upstream to the 5' UTR exon. If the upstream promoter is deleted, there is a loss of expression in the CNS. If the downstream elements are excluded, there is a loss of expression in the PNS. A further subdivision of the lower section defines regions required for expression in the PNS or antennomaxillary complex(AMX), a major collection of sensory ganglia. Both constructs that deleted a fragment just 3' of the first UTR exon showed what appears to be ectopic expression. A construct that includes the fragment does not.

Controls

The transformation vector (pFriendlyCaSper) excluding a promoter region cassette was injected as a control. One complication of this type of control is that the gene is more likely to be subject to position effects (control of an endogenous promoter after insertion). Normally, the inserted cassette serves to distance the reporter gene from any neighboring promoters. Also, the pFriendlyCaSper vector does not have a promoter, a complete coding sequence of the reporter gene (it is missing the first 43 codons), nor a canonical Kozac site. Of 13 lines stained, two showed position effects, but this was in lower gut tissues which does not correlate with the expression pattern determined by the in situ of Kv3. Additionally, the background stock (w^{1118}) was stained for endogenous expression of ß-galactosidase. None was detected; however some endogenous expression has been reported in the salivary glands of larvae. The expression pattern of the Kv3.1 reporter gene is exclusively neural and does not overlap with either the basal expression pattern of the pFriendlyCaSper vector nor the occasional pattern of endogenous b-galactosidase activity.

In our transformation studies we also made use of the pHPelican transformation vector. However, we did not inject, as a control, the pHPelican vector since this vector contains insulator elements (Suppressor of Hairy) that theoretically serve to isolate the promoter. The vector is also arranged such that the P-element integration sites are distal to the promoter with coding regions of both the reporter genes (white and lacZ) serving as buffers. This places the inserted regulatory region in the center of the cassette, as far from the host DNA as possible. This is a widely-used enhancer trap vector that has been reported to display minimal position effects and predictable leakage expression of the

heat-shock promoter in the mid-gut and lower gut.

X-gal staining was used to display the expression of the transgene. However, if the deletion constructs were to reduce the levels of expression substantially, then the sensitivity of the assay might prevent detection. Additionally, fine structures may not be distinguishable in an X-gal stain. Lastly, the enhancer trap lines are notorious for very low expression levels, even with strong promoters. Therefore, antibodies to ß-galactosidase were obtained and some of the transformant lines were re-assayed. This was particularly useful in detecting expression in the PNS and sensilla. Embryos were immuno-histochemically stained with anti-HRP to reveal the fine structure of the nervous system. Anti-HRP has a coincidental cross-reactivity to glycoproteins specific to the Drosophila neural tissues. They were also stained with anti-ßgal antibodies, and fluorophore conjugated secondary antibodies. The areas where the fluorescence is overlaid showed expression in the neural tissues. However, after a survey of more than 65 lines, it was decided that staining for ß-gal activity would be sufficient for separation of the expression controlled by the two different promoter regions and that neither would be confused with endogenous activity, leakage, or position effects.

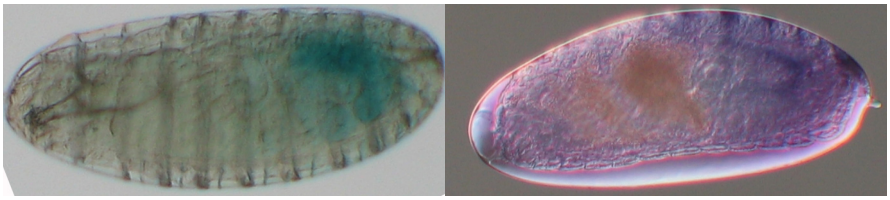


Figure 34: Endogenous ß-gal activity and transformant line with the vector alone.

Two w^{1118} embryos— stage 11 (left); and a stage 8/9- (right) that is transformed with pFriendly CaSper alone. In the embryo on the left, the staining is in the lower gut indicating endogenous ß gal activity. This occurs in some lines and not others (1 out of 13 tested lines). In the embryo to the right it is in a non-sensical pattern for neural expression that is likely a position effect wherein the transgene has come under the control of another promoter.

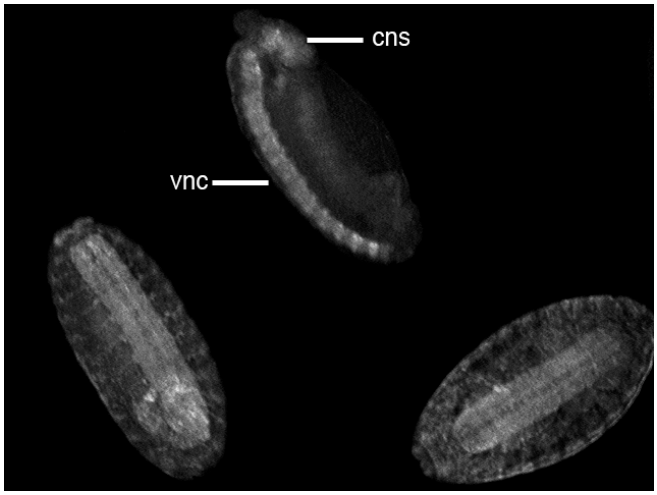


Figure 35: Anti-HRP Staining of stage 13 w1118 embryos.

Embryos that were 18-20 hours old (near stage 13) were treated with rabbit anti-HRP antibodies and goat anti-rabbit secondary antibodies conjugated to a fluorophore (TRITC). *In Drosophila*, anti-HRP recognizes an epitope found on glycoproteins specific to neural cells. The embryos are arranged in three orientations to show the structure of the nervous system from different aspects. The long structure is the ventral nerve cord (vnc) and the central nervous system (CNS) is contained in the procephalic lobes at this point. The smaller dots along the periphery are the PNS with motor-sensory neurons in a pattern similar to the larvae and adult.

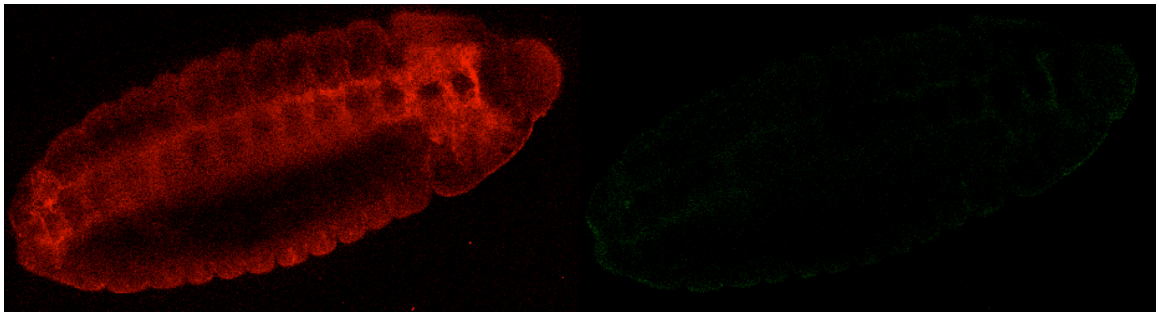
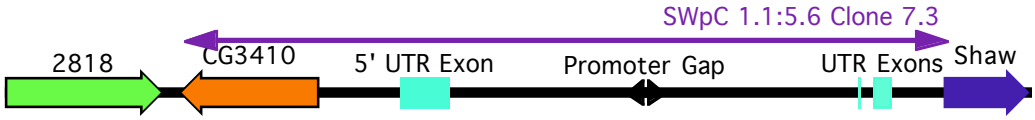


Figure 36: Anti-HRP of w1118 embryo (left) and same embryo with Anti-ß-Gal.

This served as a control for the two primary anti-bodies used to stain for neural tissue and the ß-Galactosidase transgene. Anti-HRP shows normal staining of the ventral nerve cord in the stock used for injections and the Anti-ß-Gal antibody has minimal background. There also some auto-fluorescence in the emission bands for FITC which can be seen in the gut of untreated early embryos (not shown).

Purple- SWpC1.1:5.6 (Full-length)



The full-length construct was designed to include all the genomic material from the first 14 amino acid codons of the Kv3 (Shaw) coding region 5' to the next gene on the same strand. This includes a gene on the opposing strand, the three 5' UTR exons and the first 14 amino acids of Kv3 fused to the the ß-galactosidase gene. Since no other 5' UTR exon was found upstream, it is reasonable to assume that the promoters driving native expression are contained in the construct. Since the CG3410 is not conserved between most of the fly species, it may be a recent insert into the genome. A second construct (Brown), that eliminates this gene, shows a similar pattern. Some regulatory elements have been found in the first intron before the coding exon of other genes, but there is no conservation of sequence in this intron in Kv3, and so this intron was excluded from the construct. The spatial pattern of the reporter expression is consistent with the endogenous pattern.

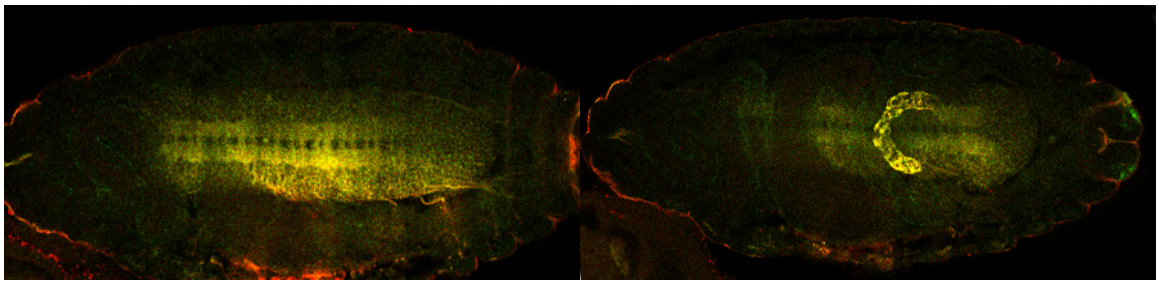


Figure 37: Double antibody staining of embryonic full-length transformant line.

Confocal image of a stage 15 embryo stained with antibodies against HRP (red) and ß-galactosidase (green). These are overlay images from both emission spectra with coincidence showing as yellow. The left image is a ventral view in a plane that is a transect of the ventral nerve cord. The right image is in a more medial plane that is includes the ring gland.

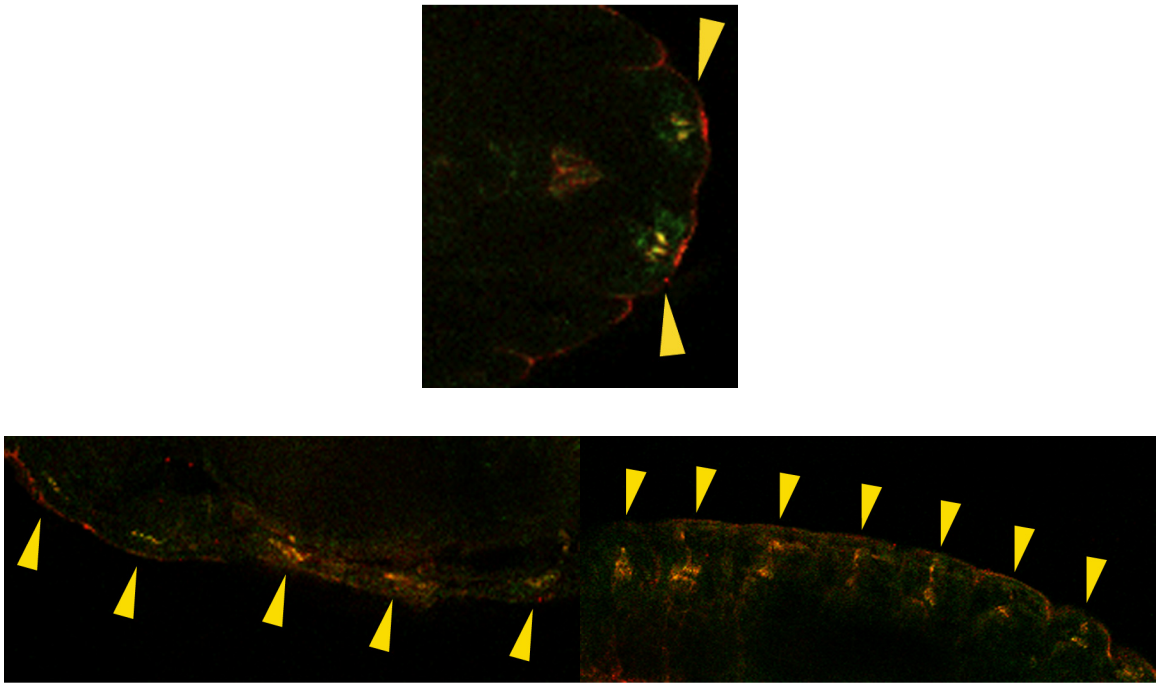
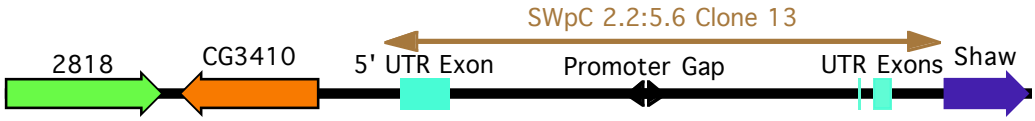


Figure 38: Details of the expression pattern of the Purple Construct

The images in the previous figure showed the general expression in the CNS. These images show the expression in the more distal portions of the nervous system. Staining of the antenno-maxillary complex is evident at the anterior end shown in the enlargement (top, yellow arrows). The green staining is non-neural tissue (not recognized by anti-HRP antibodies), possibly glial support cells. At the bottom are enlargements showing the expression in the PNS (yellow)

Brown- SWpC2.2:5.6 (Minimal-length)



The next logical construct was designed to find the minimal core promoter region that still expresses in a similar pattern to the full-length construct. We removed the genomic material coding for the CG3410 gene and possibly its promoter. The fragment begins 140 bases upstream of the transcription start site defined by the RACE reaction. The remaining material includes conserved sequences found in 9 species of fruitflies. The sequence analysis of the genomic alignments from these species hints that the regulatory sequences for this 5' promoter are downstream of the 5' UTR exon. The CG3410 gene is only found in the closest relative to *D. melanogaster*: *D. simulans*. In all other species sequenced, there is no conservation in the region between 28218 and the 5' UTR Exon.

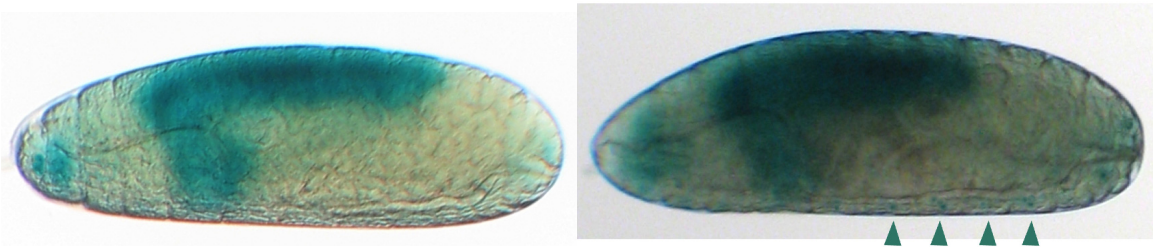
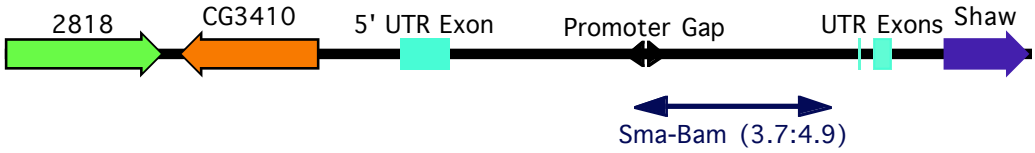


Figure 39: Minimal Control Region Construct- brown

Left: Stage 14-15 embryo. Antennomaxillary complex staining is becoming more obvious. Right: Stage 16 embryo with faint PNS staining appearing (arrows) along with dorsal caudal sensilla staining at the right. Anterior is to the left, ventral is at the top.

Blue- SWpHPel-3.7:4.9



A 1.2 Kb fragment containing the three conserved blocks downstream of the annotated promoter gap was sub-cloned into the pHPelican enhancer trap vector and used to produce 23 lines of transformed flies. Three randomly selected independent lines produced the pattern below. For each transformation, multiple lines (dozens) of flies are produced. However, due to position effects in which the promoters of the reporter construct are brought under the influence of host genetic material, there is variability in the patterns seen across all lines. For this reason, we chose to have a minimum of three lines (typically 5 or more) with similar patterns before reporting the expression of the construct. A natural development in the production of enhancer trap constructs is that some of the fragments clone in a reverse direction that is revealed upon sequencing. These were injected also to determine if the elements contained in the fragment had a specificity for a particular orientation. In most cases, the reverse fragment was different from the normal orientation. Blue transformants in which the fragment is reversed are less distinct and consistent. This may be due to several possibilities, but this might be a case of repressors or down-regulators at the 5' end of this fragment. Another scenario is that the elements in this block are orientation-specific and the recruitment of transcription factors has greater efficacy with the promoter downstream. The blocks included in this fragment are more conserved (≥ 85%) across all Drosophilid species tested than the UTR exons. The core promoter elements contained in or near those exons appear to be less functionally constrained than these regulatory elements. The fact that these elements

drive expression of the reporter gene from an exogenous gene suggests that the specificity for expression lay at the enhancers and repressors, not in the core promoter regions. This fragment contains the most conserved of all the putative regulatory elements. Based on the Orange constructs (described below), these elements appear to act on the downstream promoter (a TATA box promoter) to drive expression in the PNS.

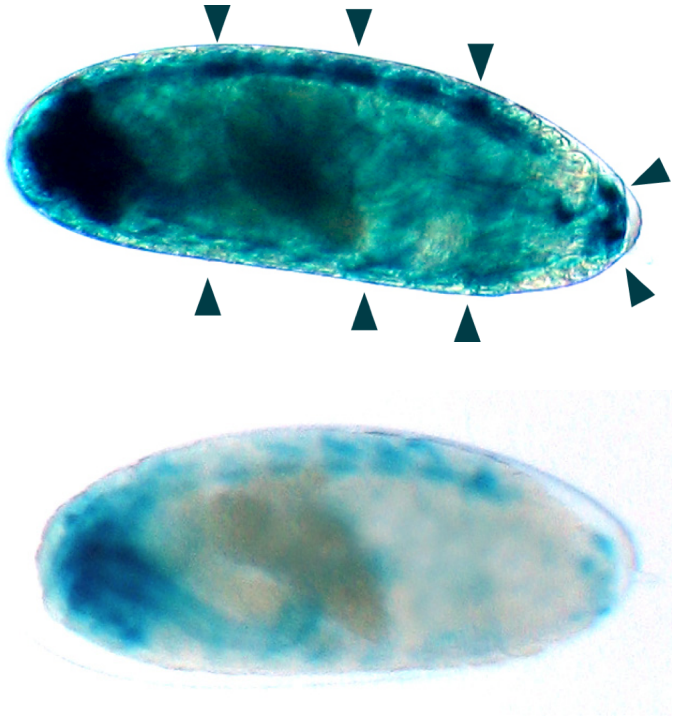
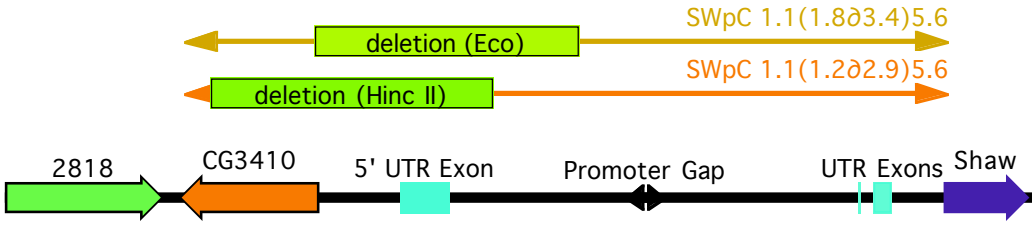


Figure 40: Expression of ß-galactosidase with the Blue enhancer-trap construct.

Top: X-gal staining shows strong staining in the antenno-maxillary complex and the peripheral nervous system (arrows). This is a nearly-ventral view (the embryo is rolled slightly on its side) with anterior to the right. The staining towards the posterior is likely endogenous with some collection of X-gal substrate in the trachea. Embryos must be strongly stained to reveal the details of the PNS.

Bottom: Better sample showing the detail of the PNS, but resulting in fainter staining of the AMX. Image was post-processed by lightening the gamma.

Oranges-SWpC1.1(1.8∂3.4)5.6 and SWpC 1.1(1.2∂2.9)5.6



In these two constructs the entire upstream promoter and some of the surrounding sequences have been deleted from the Purple full-length construct. The light orange construct (with an EcoR I fragment deleted), removes more of the conserved blocks between the Promoter Gap and the 5' UTR Exon than the dark orange lines. The dark orange construct (with Hinc II fragment removed), removes the CG3410 gene and the 5' UTR exon, but leaves the conserved blocks in the control region intact. Flies from the two different transformant groups are similar, with the primary distinction being that there is no expression in the ventral nerve cord and procephalic lobes of the CNS. This argues that either the upstream promoter is necessary for expression in the CNS, or that the elements just 3' of the UTR exon must be present for CNS expression. There is a great deal of what appears to be ectopic expression. This may be the result of the loss of a regulatory element immediately down stream of the exon. Another possibility is that elements that normally drive expression from the upstream promoter are now enhancing expression from the downstream promoter, which may be causing miss-expression of the reporter gene. A series of new constructs have been produced that will examine this deleted region in more detail.

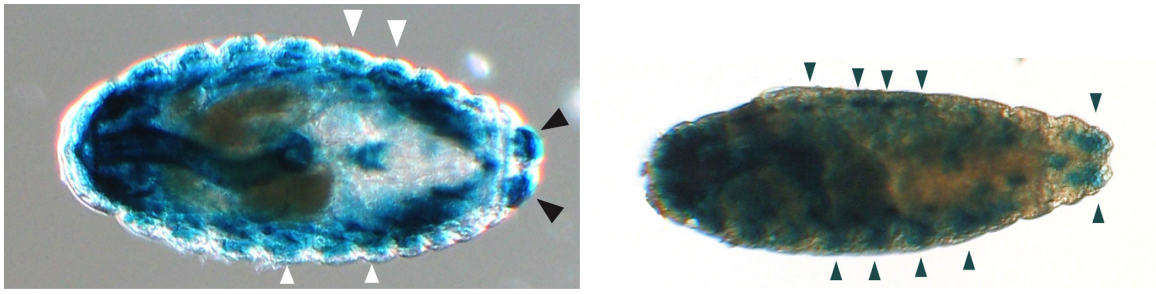


Figure 41: Upstream promoter deletion construct expression.

The light orange construct expression pattern is shown to the left (grey background) and the dark orange to the right. These are ventral views with anterior to the right. Both are heavily stained (over night) with X-gal to reveal PNS pattern expression (arrows). There is no staining of the ventral nerve cord or procephalic lobes.

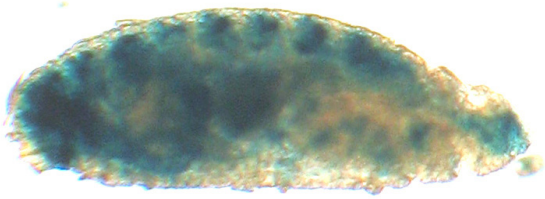
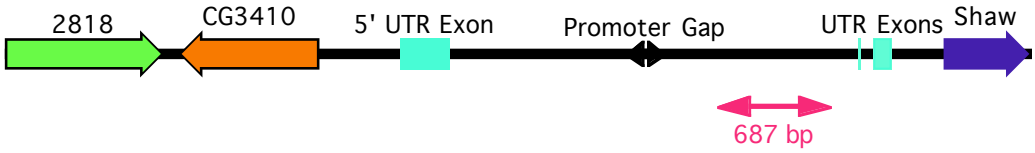


Figure 42: Dark Orange construct

Another example of Dark Orange construct staining that reveal the “twin spot” peripheral nervous system staining along the top. There is still a great deal of ectopic staining. The image was post-processed by lightening the gamma universally across the image.

Pink-SWpHPel-4.2:4.9



The Pink construct inserts the 687 bp fragment that contains two conserved blocks into the pHPelican vector. Ten lines were produced with two having the block in reverse orientation. These blocks were isolated as a subset of the Blue and Orange constructs to determine their function independent of the next upstream block.

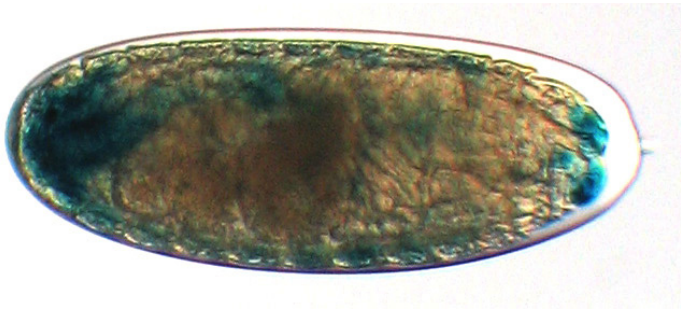
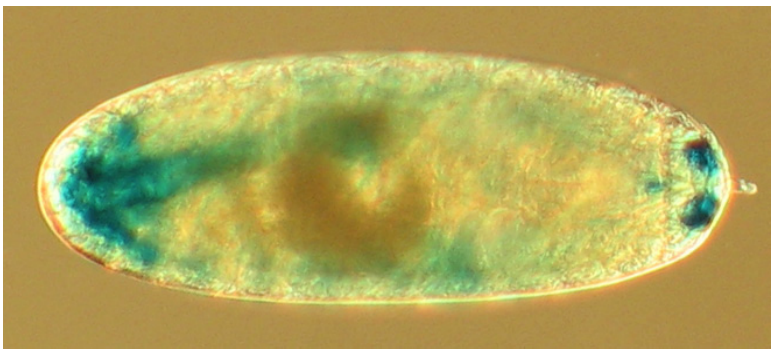


Figure 43: Pink construct expression pattern.

Ventral view with anterior to the right. Staining is strongest in the antenno-maxillary complex seen at the anterior end. There is faint PNS staining, but not as heavy as with the Blue or Orange constructs that include the third (and most conserved) block. In the image below, the fragment is in reverse orientation.



Aqua-SWpC-1.1(4.2∂4.9)5.6

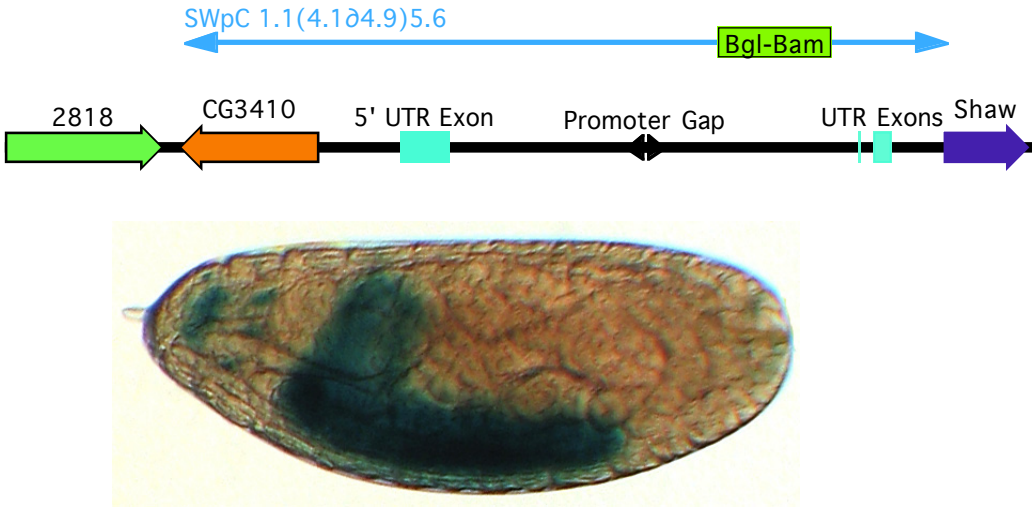


Figure 44: The aqua construct shows expression largely confined to the CNS.

X-gal staining of stage 14-15 embryo. Anterior is to the left. The ventral nerve cord is at the bottom with clear staining of the procephalic lobes and some staining of the labial members of the antennomaxillary complex.

The aqua construct removes only the 687 bp fragment that is expressed in the Pink construct. The remainder of the Purple (full-length) construct is intact. While the PNS and midgut staining that appears to be driven in the Pink construct is largely absent, there is some staining in the antennomaxillary complex. This may be due to the presence of some elements that are just upstream of the deletion that are responsible for expression in various parts of the complex. In the absence of the elements normally found in the region deleted, there is some expression but not as strong as in the full-length reporter construct.

Olive-SWpHPel-2.2:3.7

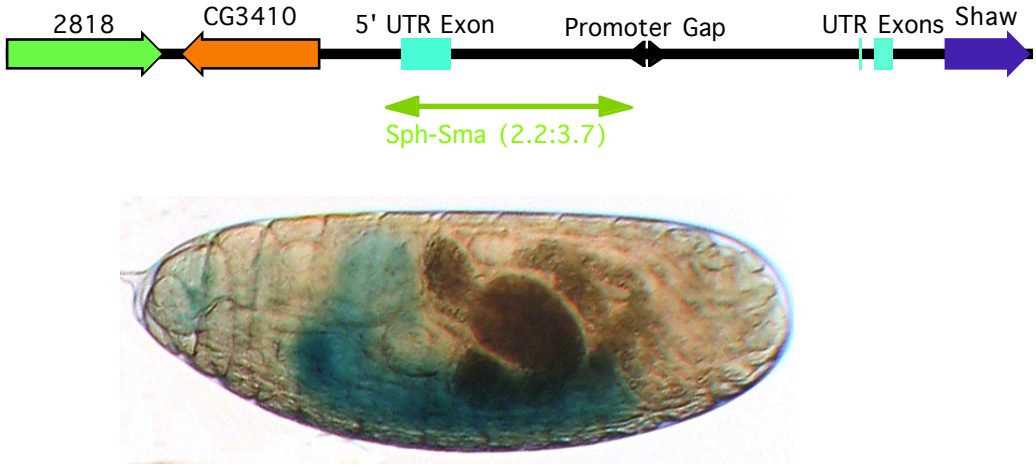


Figure 45: The olive expression in the CNS.

This transgene takes the fragment that includes the upstream promoter and the majority of the conserved blocks upstream of the promoter gap. Expression is largely limited to the CNS. The staining is not as strong as in the equivalent deletion in the purple construct. This may be due to the competition of the promoter elements between the two basal promoters present— one in the included 5' UTR exon and the other in the vector itself that drives expression of the reporter gene.

The olive construct is essentially the inverse of the light orange construct. The 1.2 Kb fragment that was deleted is now placed in the pHPelican vector. There is a specific element just downstream of the Promoter Gap missing in this fragment and this appears to weaken expression generally, and all but loses expression in the PNS and antennomaxillary complex. Some of the reduction of expression may be due to the presence of two promoters, the native core promoter in the 5' UTR exon, and the heat shock promoter in the vector. There is possibly promoter competition occurring. More constructs that capture only fragments down stream of the exon were built, but not successfully injected.

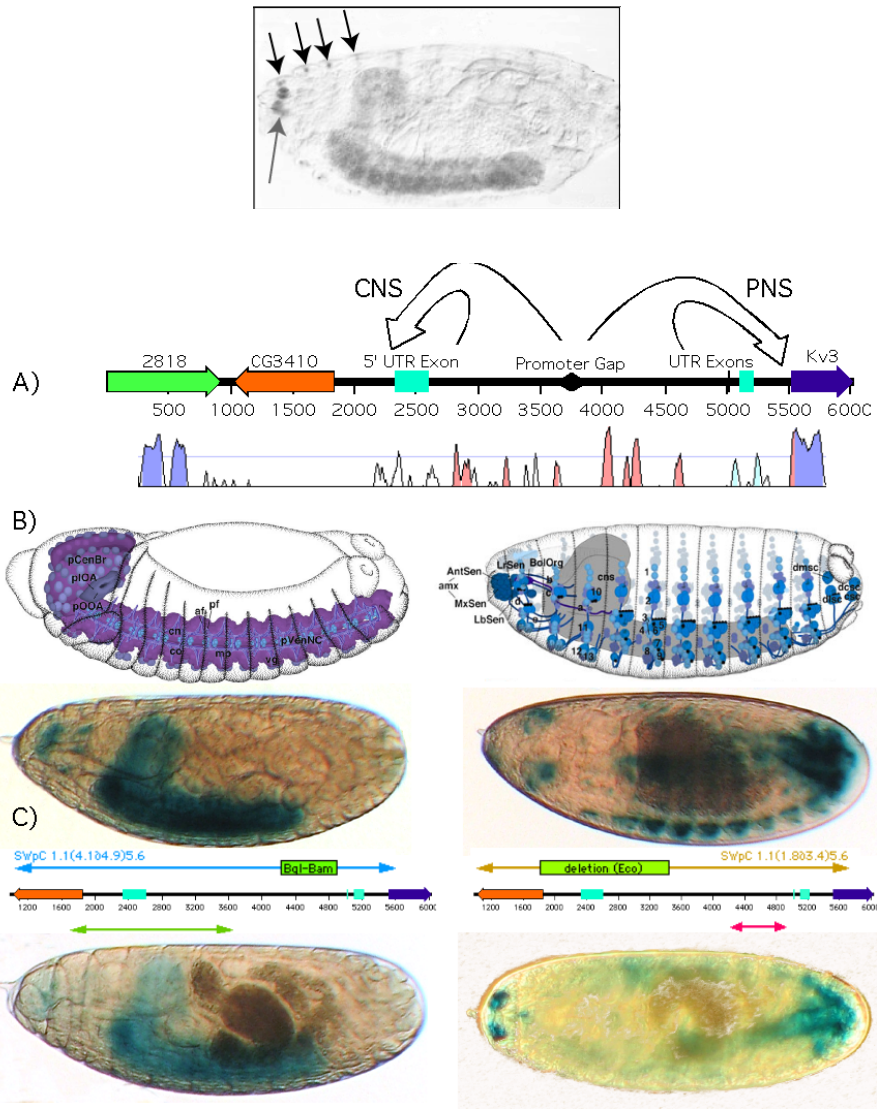


Figure 46: Summary of expression data.

At the top is the embryo *in situ* showing expression in the CNS, PNS, antennomaxillary complex, and dorsal causal sensilla. In panel A, a map of the regulatory region annotated with the RACE results is aligned with a Vista plot showing conserved blocks in the *Drosophilids*. Panel B is a schematic of the CNS structures (purple) on the left, and on the right, the PNS is shown (blue) along with the antennomaxillary complex (labeled). Panel C has the four key expression constructs dissecting the two promoter regions into CNS (left) and PNS (right) expression. The two images above the construct maps are deletions from the full-length "purple" line, while those below show the expression of the enhancer trap constructs with the fragments from the deletions inserted.

CONCLUSIONS

Chapter 6: Identification of Neural Enhancers

IN SUMMARY

No single approach to the study of the transcriptional regulation of the Kv3 gene would elicit the information found here. The design of the constructs used for the *in vivo* functional analysis required information from the *in vitro* RACE and the *in silico* alignments and sequence analysis. Garnering sequence for the *in silico* analysis required the in vitro mapping of the promoters and knowledge of the expression patterns of the gene for comparison to similar genes, promoters and elements. It is through the integration of all three approaches that the regulatory region is described.

in vitro

The transcription start sites for four *Shaker* cognate genes were identified. Each gene has two different promoters; one that begins transcription at the first coding exon and another further upstream. In each case, two or more 5' UTR exons were found upstream. In Kv2, the first exon is 30 Kb 5' of the coding exons, with one intervening gene on the opposite strand. The first coding exon of Kv4 was extended 352 bases in the 5' direction and a new 5' transcription start site identified 10 Kb upstream. Transcripts from the two specific promoters showed a much greater rate of transcription from the 5' promoter in the embryo. Two promoters, a TATA-less promoter 6 kb 5', and a TATA promoter at the coding exon, were identified for Kv3. Two more UTR exons were found between the promoters. A fourth member of the group, Kv3.2 was identified from homology searches in the Anopheles genome and traced back to the Drosophila genome.

It is designated 3.2 due to its similarity to Kv3. This gene also has two promoters with the upstream promoter 19 Kb away. There are two intervening genes on the opposite strand.

The genomic region upstream and including a portion of the next 5' gene above Kv3 was cloned from *D. melanogaster* and used as the starting material for a series of transformation constructs designed as a functional analysis (*in vivo*) of the regulatory region. The extent of the deletions and which fragments to include in this analysis was determined from sequence analysis (*in voltro*).

in voltro

Sequence analysis was performed on the discrete *D. melanogaster* sequence along with homologous fragments isolated from six other species of fruit flies. The generation of a multi-species alignment revealed regions of highly conserved non-coding sequence. These were more conserved than UTR exons and core promoter sites. The gene annotated 5' of Kv3 was not found in most other species of Drosophila. This appears to be a recent inclusion in this region. A Bayesian search for core promoters showed that the suspected 5' transcription start is a strong candidate under stringent search conditions. A region in the center of the control region shows near zero conservation, but is not typical of matrix associated sites. It is less conserved than intronic regions within the gene. It was designated the 'promoter gap'. Three specific highly conserved blocks were hallmarked for functional analysis. The region from the 5' UTR exon to the promoter gap was cloned or deleted in constructs for functional analysis. Meta-MEME searches found Inr-DPE (type 10) segments in the upstream promoter. MEME also identified a recurring motif with extremely low expect values in the upper control region (5' of the promoter gap). A Transfac search of these motifs failed to show any particular match to

known transcription factor binding sites.

in vivo

The native expression pattern for Kv4, Kv3 and Kv3.2 was determined with embryo *in situ* hybridization. All three were found in the early stage 13 neural tissues, specifically the ventral nerve cord and the procephalic lobes.

By stage 15, expression in the peripheral nervous system was detectable in the antennomaxillary complex and the dorsal caudal and lateral sensilla.

Eight transformation construct lines were eventually tested in flies. More than 20 were produced, but one particular vector line was determined unsuitable for these purposes.

A full length reporter, the so-called purple line, recapitulates the expression pattern of Kv3 found with the *in situ* assay.

A second construct, brown, removes the CG3410 gene upstream of the 5' UTR and maintains the native expression pattern. This argues that all of the regulatory elements responsible for expression of Kv3 can be found between the two promoters. This correlates with the conserved sequences found in the multi-species alignment.

The orange constructs remove the upstream promoter to reveal expression in the peripheral nervous system and a loss of expression in the CNS. These constructs leave some of the conserved regions (and those identified as repeating motifs) intact. There is a large amount of ectopic expression in muscle and gut tissue. This could be an example of unregulated expression at an alternative promoter, or that a discrete element deleted in both constructs is responsible for suppressing such ectopic expression.

To test the remaining fragments of the control region, two enhancer trap constructs were built with fragments down stream of the promoter gap that includes the three conserved peaks. The blue construct shows strong expression in the PNS and AMX with some mild ectopic (or enhanced endogenous) expression in the midgut.

The pink construct has a much more defined pattern in the AMX with faint expression in the PNS.

The aqua construct deletes the pink region and PNS expression is all but eliminated. There is faint expression in the AMX.

The olive construct is solely the first 1.5 Kb fragment from the brown construct and it shows expression solely in the CNS.

in toto

The data from *in vitro*, *in voltro* and *in vivo* studies were necessary to define the regulatory elements that control the expression of Kv3. In each line of this research, difficulties required some form of optimization of established techniques (along with some novel tactics) to extract the information about the transcriptional control of just this one gene. It would seem that while whole-genome approaches are being proposed for the study of gene regulation, some tried and true grunt work will always be required to this end.

Materials and Methods

Many of the techniques used in this work are based upon standard protocols. However, for the purposes here, most were modified to accommodate particular challenges. Any who follow this path will likely find a different set of hurdles to overcome. They are encouraged to be creative; most of the results above arise from sheer stubbornness and taking many different approaches.

Chapter 7: RNA and DNA

RNA PREPARATION

Total RNA

Modified RNA preparation techniques based on the One-Step method (Sambrook) were used. Drosophila represent a challenge in that the cuticle mass makes separation of proteins from nucleotide material critical. To this end, often two or three cycles of acid-phenol extraction and subsequent precipitations proved effective.

The general protocol.
Reagents:
Guanidine isothiocyanate (Sigma)
1M sodium citrate, pH 7 (DEPC-treated, autoclaved)
Sarcosyl (N-lauryl sarcosine, Sigma)
b-mercaptoethanol (Sigma)
2M sodium acetate, pH4 (DEPC-treated, autoclaved)
3M sodium acetate, 100mM magnesium acetate, pH 5.2
Absolute ethanol
Propan-2-ol (isopropanol)
70% ethanol (made with DEPC-treated, autoclaved water)
0.5% SDS (made with sterile, DEPC-treated water)
Denaturing Solution- 4M guanidine isothiocyanate, 25mM sodium citrate, pH7, 0.5% sarcosyl, 100mM ß-mercaptoethanol.
(Denat. Sol. can be made and stored at 4°C without ß-mercaptoethanol for several months. ß-mercaptoethanol should be added to 100mM immediately prior to use.)

1) Tissue is homogenized as rapidly as possible, at 4°C, in solution D (500ul per 50mg tissue) with an eppendorf pestle homogenizer until a smooth, lysed, homogenous suspension is obtained.
2) Add 50ul 2M sodium acetate, pH4.0 and mix vigorously.
3) Add 500ul phenol and mix vigorously.
4) Add 100ul chloroform, mix vigorously and incubate on ice for 15 minutes.
5) Centrifuge mixture at 10,000g for 10 minutes in a microfuge at 4°C.
6) Remove upper, aqueous phase to a clean, sterile, DEPC-treated eppendorf tube. After centrifugation, RNA is present in the aqueous phase while, due to protonation at the acidic pH used, genomic DNA is partitioned into the phenol phase.
7) Extract the upper aqueous layer with an equal volume phenol/chloroform and centrifuge as before. Repeat the extractions until no interface material is seen.
8) Precipitate the aqueous phase by the addition of an equal volume (500ul) of propan-2-ol. Incubate at -20°C for 20 minutes.
9) Pellet RNA by centrifugation at maximum speed in a microfuge for 10 minutes.
10) Wash the RNA once in 70% ethanol and vacuum dry.
11) Re-dissolve in 200ul 0.5% SDS at 65°C.
12) Extract with an equal volume (200ul) of phenol/chloroform as above. Repeat until no interface material is visible.
13) Precipitate pure RNA by the addition of 20ul 3M sodium acetate, 100mM acetate, pH 5.2 and 500ul absolute ethanol. Incubate at -20°C for 20 minutes.
14) Pellet RNA by centrifugation at maximum speed in a microfuge for 10 minutes.
15) Wash the RNA once in 70% ethanol and vacuum dry.
16) Dissolve RNA in appropriate buffer i.e. DEPC-treated, sterile TE, pH 8 or 0.5% SDS if no enzymic manipulation of the RNA is needed. SDS is an inhibitor of ribonucleases.

Total RNA was sufficient for most RACE, probe and cDNA preparations. If not, poly-A purification and cap-selection were used to increase the amount of full-length message in the samples.

Poly-A RNA

Poly-A RNA was purified from 10 μg aliquots using oligo-dT Sephadex (Oligodex) columns and standard techniques (Sambrook). Final yield was typically less than 100 ng per sample.

The general protocol.
Before using oligo-dT bound column material, it is best to wash the binding medium to remove fine particle of latex. To wash the latex, transfer appropriate amount of beads/ media (300 ul of the suspension of Oligotex per 1mg of total RNA) into a microfuge tube. Spin for 3 min at 12,000 rpm. Discard the supernatant. Gently suspend the latex in the same volume of 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.1% SDS. Spin again. Gently suspend the latex in the same buffer.

1. Add 1 mg of total RNA dissolved in RNase-free water to 300 ul of the Oligotex-dT30 suspension.
2. Incubate for 3 minutes at 65 C. Chill on ice.
3. Add 0.2 volume of 5M NaCl. Incubate for 10 minutes at 37 C.
4. Centrifuge for 3 minutes at 15,000 rpm. Discard the supernatant.
5. Suspend the pellet in 1 ml of washing buffer (10mM Tris-HCl pH7.5, 1mM EDTA, 0.5M NaCl, 0.1% SDS).
6. Centrifuge for 3 minutes at 15,000 rpm. Discard the supernatant.
7. Suspend the pellet in 300 ul of RNase-free water containing 0.1% SDS.
8. Incubate for 5 minutes at 65 C. Chill on ice.
9. Centrifuge for 3 minutes at 15,000 rpm. Transfer the supernatant into new microfuge tube.
10. Carry out phenol-chloroform extraction and ethanol precipitation by standard procedure. Rinse the pellet with 75% ethanol. Dissolve the poly(A)+ RNA in 10 ul of RNase-free water.

Cap-selection of RNA and 5' RACE

Here Calf Intestinal Phosphatase (CIP) is used to remove 5' phosphates from uncapped (either unspliced or degraded) RNA. A subsequent Tobacco Acid Phosphatase (TAP) reaction then cleaves the 7-Methyl-Guanidine capped (full-length) messages. This now leaves a 5' mono-phosphate. The RNA linker is then ligated to the RNA pool. The adaptor will only ligate to the full-length message which has the phosphate necessary for the reaction. Now an RT-PCR reaction is performed using thermo-labile high fidelity reverse transcriptase (Superscript III, Invitrogen) and primers specific to the known 5' end of the message and the RNA linker.

The general protocol:
1. 500 ng poly(A) RNA is used in a CIP reaction with 2 μl buffer, 2 μl CIP (10 u) and water to 20 μl. 37° for one hour.
2. Add 15 μl 3M Ammonium Acetate pH 5.2, 115 μl dwater, 150 μl acid phenol:chloroform. Vortex, centrifuge 5 min. 13k g, extract top layer, add 150 μl chloroform, spin, 5 min 13k g, extract top layer, add 150 μl isopropanol, chill to precipitate, centrifuge 20 min. 13k g at room temperature, rinse with 500 μl chilled 70% ethanol, centrifuge 5 min. 13K g. Re-suspend in 8 μl 1x TAP buffer.
3. TAP reaction: 8 μl RNA-in TAP buffer, 2μ TAP. 37°, 1 hour.
4. Ligation reaction: 2μl CIP-TAP treated RNA, 1μl RNA Adaptor, 1μl buffer, 2 μl T4

RNA ligase, 4μl dWater. 37° 1 hour.
RNA Adaptor-
5'GGGUUCGGGCUUAGGCUCCAGUGCCUGUUCGGUGGUCGCGGCGCUGAUG
GCGAUGAAUGAACACUGCGGCAAGCCGCUUAAUGACACUCGUUUGCUGGC
UUUGAUGGGCGAGCUGGAAGGCCGUAUCUCCGGCAGCAUUCAUUCAUUUA
CGACAAA-3'
5. Reverse Transcription Reaction: 2μ ligation reaction, 4μl dNTP mix (10μM each), 10 μM gene specific primer (or Random Decamers), 2μl 10x buffer, 1μ RNaseOut RNase inhibitor, 1μl Superscript III (or Thermoscript) in a 20 μ reaction.
We incubated as high as 55°C with a prior heat denaturation of the ligation product at 70°C. This was necessary to minimize secondary structure that inhibits the progress of the reaction.
6. Nested PCR: two pairs of primers were designed (actually multiple sets were required since the adaptor and inner gene primers are extended with a BamH I cut site) for cloning purposes, which makes primer design difficult to predict. It was later discovered that shorter primers without cuts sites were more effective. The PCR reaction were conducted with conditions suitable for the primers, but typically, a range of annealing temperatures and Mg ion concentrations were used. In all cases, 1μl of RT reaction was used.
NUP Primer 5'-AAGCAGTGGTAACAACGCAGAAGAGT-3'
A dilution of 1:100 of the first PCR reaction is used as a template for the second nested reaction.
TAP-minus RNA pool was also used in a ligation-RT-PCR series as a control to ensure that products of the nested PCR were from capped message only.

Gene-specific primers:
Kv2
ShabRace-RT 5'-TACTCCAGGTCATCAC-3'
ShabEx1 5'- GGAGGAAAGCCATTGGAAACCAGGA-3'
ShabEx2 5'- GCAGCATCAGCAGCAGCAACAGGA - 3'

Kv4:
SmartOuter 5'- CGATGAAGAAACCCGTAACAT-3'
SmartInner 5'- TAGCAGTCGCCAATGACATCC-3'

DNA PREPARATION

Standard DNA preparation techniques were employed along with some modifications to suite the needs of certain procedures. Drosophila present a particular difficulty in that the proportion of protein to nucleotides is sufficiently great to necessitate additional steps in the process. In fact, several different methods were used depending upon the species.

Genomic

Welcome Bender-Jay Hirsch-Nigel Atkinson Method

Flies frozen in liquid nitrogen are ground in a ceramic mortar at -70 C and the resulting powder is added to a 2 ml Dounce homogenizer with 1 ml of homogenizing solution.

Fly homogenizing buffer:
0.1M NaCl
0.2M Sucrose
0.01 Na_2EDTA
0.03M Trisbase pH8.0

Transfer to autoclaved 15 ml corex tubes. 65°C 30 minutes
add 300 μl 8M KOAc, mix well, 0° C for 60 minutes
Spin 8K 8 minutes at 4° C (SS34 or equivalent)
Pour supernatant into 2 microfuge tubes (1.5 ml)
Spin 10 minutes microfuge at 4° C
Keep supernatant and divide into 3 microfuge tubes
Add 700 μl of room temperature 100% ethanol to each tube
Let stand 5 minutes at room temperature; then 5 minutes spin in 4° C microfuge
Wash pellet with cold 70% ethanol and re-suspend in 100 μl TE, pool all 3 tubes
Add 6 μl of 5M NaCl (final conc is 0.1 mM) and 15.6 μl of 200 mM spermine 4HCL
Spin 10 min, 13K g; remove supernatant
Re-suspend pellet in 180 μl of H20
Add 20 μl (1/10th volume) 3M NaOAc pH 5.5 and 500 μl ethanol
-20 ˚C for 30 minutes to o/n; spin 10 minutes 13K g
Wash pellet 2 times with cold ethanol

Atikinson's combined chemical lysis-double phenol-Qiagen prep

This preparation was used for large flies (for example, *D. virilis)* or *D. willistoni* which seem to defy all matters of normalcy.

Homogenize flies as above, then suspend powder into 30 mls of rapidly stirring ice-cold NIB. Transfer to ice-cold 30 ml corex tube. Spin briefly at low speed to pellet large mass material.
Transfer to a new ice-cold corex tube and spin 7000 rpm 4°C 7.5 minutes.
Re-suspend pellet in 19 ml ice-cold NIB and transfer to 50 ml conical tube. Add 8 ml 10% Sarkosyl. Mix slowly by inverting. Let stand on ice 10 minutes.
This is the second departure point, normally this would lead to a CsCl prep, but we precipitated the chromatin and brought that material up in P1 of the Qiagen prep.

Nuclear Isolation Buffer:
37.5 mM Tris [pH 8.5], 0.05 mM spermine, 0.125 mM spermidine, 0.5 mM EDTA [pH 7.4], 20 mM KCl, 0.5% thiodiglycol, 0.05% Empigen BB, 0.1 mM PMSF, (2 μg of aprotinin per ml)- the pH is different from Bingham, Levis and Rubin.

Plasmid

With the advent of Qiagen kits (Qiagen, Venlo, Netherlands) alkaline lysis preps of large or medium scale plasmids have been relegated to the past. However, for extremely clean preparation of large quantities of DNA suitable for injections, we found that the alkaline-lysis prep below produced the best results:

Alkaline lysis 50 ml midi-prep

Spin down a 16-20 hour 50 ml culture, 5K g 30 minutes.

Add 3.0 ml Solution 1; vortex to suspend pellet; let stand on ice 10 minutes
Add 6.0 ml Solution 2; do NOT vortex- mix by inverting slowly; let stand on ice, 10 min.
Add 4.5 ml Solution 3; mix by inverting slowly; let stand on ice 10 minutes
Spin, 5K g 30 min.
Pour through Kimwipe into a fresh 50 ml conical.
Add 11 ml isopropanol; precipitate pellet (30' or overnight)
Re-suspend pellet in 200 μl TE; add 200 μl 5M LiCl; let stand on ice 10 minutes
Spin, 13K 10 minutes, draw off supernatant and transfer to a clean 1.5 ml tube
Precipitate with 400 μl isopropanol (30 min. or overnight), spin 13K g 10 min.
Re-suspend in 500 μl TE and treat 10 min. with 4 μl RNase (10 μg/μl) 37° C, 15 min.
Phenol-chloroform extract, precipitate o/n with NH_3oAc and EtOH.
Yields 100-400 μg DNA

Solution 1: 1% glucose, 25 mM Tris-HCl, pH 7.5, 10 mM EDTA
Solution 2: 0.2 N NaOH, 1% SDS
Solution 3: 3 M KoAc

Cosmid/PAC/BAC prep

This is adapted from Roe (Univ. of Wash), which is derived from Chen, Pan, and Ying at the Berkeley Genome Sequencing Centre. This employs a double acetate precipitation. These are low copy number vectors, and so require staged growth:

Single colonies are streaked and grown overnight. This smear is then used to incubate a 50 ml culture that is grown for 8-10 hours. This culture is then transferred to a 250 ml flask with an equal volume of the same medium. This is grown over night for 12-16 hours and 50 mls is used to inoculate a 1 L culture in a 3 L flask. This is divided into 500 ml bottles and spun down. If pellets are frozen overnight at -70° C, there is a higher yield.

cDNA library screening

cDNA library screening is rapidly becoming a thing of the past. New, cap-selected libraries are available through commercial sources and often are *gratis* to educational institutions. In short, don't do what I did, it isn't worth it.

One side note is that once the library was plated and the screening process was begun with a radio-labeled probe generated from conserved sequence in the pore region, cDNA clones were obtained for all three genes. None was of any true utility, as a RACE reaction is far more informative. Eventually, this library was mishandled by persons known only to themselves and it was killed, rendering all screens useless.

Genomic cloning: RAGE, Vectorette, Degenerate Primer PCR

In order to clone genomic material surrounding the genes of interest from related species, three different techniques were employed. RAGE applies principles from RACE, hence the name: Rapid Amplification of Genomic Ends. This is useful when smaller fragments are produced after a restriction digest of genomic material. Should the fragment exceed a practical size for PCR, Vectorette is employed. Lastly, if there is an upstream gene a reasonable (4-10 Kb) distance upstream from the target gene, it was sometimes possible to use degenerate primers to that gene's coding sequence.

RAGE

In this modified RAGE technique, an anchor sequence was produced from fish genomic material (Sternopygus) with a pair of primers that generated a 5'-CTAG-3'

overhang after digestion with BsmF I. This enzyme was chosen since the fish-anchor had the CTAG sequence near the end, and BsmF I has a particular tolerance to digest at the end of fragments and would produce the overhang without a palindromic site. The primers were chosen to amplify a sodium channel with no homologue in Drosophila. The resulting sequence was also used in a BLAST search against the NCBI genome database. This particular overhang was chosen as it has the exclusive ability to anneal with the sticky ends from genomic fragments digested with four different restriction enzymes: Xba, Spe, Nhe, and Avr II. This increased the likelihood that a genomic restriction fragment would be produced of a size suitable for PCR. Often a region of interest will be devoid of two or three of the cutters, but rarely all four.

Genomic material from 6 species of Drosophila (sechellia, virilis, immigrans, pseudoobscura, mulleri, and hydei) was prepared and digested with the four enzymes. The genomic material from D. willistoni was also digested, but all attempts at PCR failed. The fish-anchor was ligated overnight at 16° C, and PCR performed with standard protocols over a 20 degree range of annealing temperatures.

Vectorette

Vectorette is similar to the RAGE technique, however in this instance high-capacity vectors (BACs), were employed to clone and produce mini libraries. Single colony copies were pooled (100-300 from a plate) and PCR performed on boiling preps of the mélange. If a product was produced from primers designed to the conserved coding region, then that pool was further subdivided until individual colonies could be tested. Once a colony carrying the desired insert was localized, then the size of the insert was determined and a series of walking sequencing reactions preformed to obtain

upstream sequence.

Degenerate primer PCR

Potassium channels in Drosophila have an N-terminal tetramerization (T1) domain that is highly conserved. There is also a canonical pore sequence (GYGD) along with more conserved amino acids in the pore loop. These made cloning of the coding region of the genes from different species relatively straightforward. The difficulty is in designing degenerate primers to the 3' termini of the next 5' gene. This is typically UTR and therefore likely to be quite variable.

For cloning of Kv4 genomic, degenerate primers were designed to the 5' most coding exon where the T1 domain is and to the pore region. There were many (67) primers used but these are the critical sets. Much of this has been rendered obsolete with the sequencing of many *Drosophilids*:

NewShal-InnerL: 5'- GGGAATTTGCGGGGGATTTA - 3'
NewShal-InnerR: 5'- CAGCCAATCGTCTGTAAACTG - 3'
NewExon0: 5'- GGACAG/CATG/CGAAGAC/GCAT/ACC - 3'
NewExon1: 5'- GCATG/CAAG/CAA/CCCCGTAAC/GATA - 3'

For cloning of Kv3 genomic degenerate primers were designed to CG3410. However, it was eventually discovered that this gene is not extant in the other species of flies used. So a second set of primers were designed for CG2818 and the first, second, and third coding exon of Kv3.

CG3410: 5'- GGAAACGTCTGCTTTCAAGG-3'
2818Ex3: 5' – CCTACGAAAGGCAGGGAATC – 3'
2818Ex2: 5' - ATGTCCTGACCACCAAGGAG -3
ShawEx1: 5' - AATTCCTCTTCGGACGGTTT -3
ShawEx2: 5' - GCTCCACTTGGTTCGAGTCT -3
ShawEx3: 5' - TTGAGCACTTGTGCCAAGAC -3'

This last primer has a single base change that differs from genomic. It should be:
ShawEx3: 5' - TTGAGCACTTGTGCGAAGAC -3'

The fragment produced from the 2818Ex3::ShawEx3 is 6039 bp. This is the core fragment for the production of all sub-clones. This fragment was cloned into PCR TOPO 2.1.

SUB-CLONING AND CONSTRUCTS

Sub-cloning of genomic fragments in the putative regulatory region typically took the course of PCR with matched primers across known useful restriction sites, or with restriction enzyme cut-sites built into the primers. These fragments were assembled in a standard sub-cloning vector such as TOPO PCR2.1 (Invitrogen, Carlsbad, Ca) or pBluescript (Stratagene, La Jolla, Ca); transformed into XL1-blue or Top10 DH5α cells, and DNA prepared with standard or in-house plasmid prep techniques. The fragments were then digested from the vector; gel-purified and ligated into one of several tranformation vectors: pPTGAL, pUASP, pHPelican, pPelican, pCaSper-AUG-ßGal, and FriendlyCaSper. This last vector is a modified version of pCaSper with a new polylinker inserted and the first 33 amino acids removed from the lacZ reporter gene to allow generation of a fusion protein. Many of the constructs were either abandoned or determined to be ineffective.

Kv4

All of the promoter region fragments were subcloned into TOPO-pCR2.1 and then shuttled to the pPTGAL vector, which proved to be an ineffective transformation vector for this region. More than 24 clones were generated in fragments and then ligated to produce larger fragments. Any future cloning scheme would require a different approach dependent upon the needs of the transformation vector. A long discourse on the cloning path of Kv4 is omitted to protect the innocent.

Kv3

Purple

A full length construct (Purple) was produced by PCR from the genomic clone (1.24). The 2818Ex3 was used with new primers:

124aa3-Mfe: 5'- ATCCAATTGATTCATGGCTGCATGCGCATA - 3'
124aa14-Mfe: 5'- ATCCAATTGCACCACCCTGTTTTCCGAGTC -3'

The product has an introduced Mfe I site, which once cut, anneals into an EcoR I cut pFriendlyCaSPer. This puts the first 3 or 14 amino acids in-frame with the coding sequence of the lacZ reporter gene. Clone 7.3 was sequenced and confirmed and then injected as the first line of full-length flies (purple).

Figure 47: The Purple Cloning Scheme

Two full-length fragments were cloned such that the first few amino acids of Kv3 were in frame with the coding region of the reporter gene in pFriendlyCaSper. This is a modified Lipschitz vector with a gutted multiple cloning site and 43 amino acids of lacZ removed.

Oranges

The light orange construct was built by modifying the purple. Purple was digested with EcoR I and then ligated shut. The resulting ligation was transformed and picks screened for correct sequence. The dark orange contruct required a different approach. The original genomic clone, 1.24, was digested with Hinc II and then ligated shut. The same primers that produced purple were used for a PCR reaction on the gutted genomic clone. The final product was generated in the same fashion as purple.

Aqua

This removes a fragment from purple with a Bgl II BamH I digest. However, a second BamH I site exists in the CG3410 gene which must first be removed with a Pst I digest. Purple is digested with Pst I, ligated and transformed. Then a double digest is performed and ligated shut.

Brown

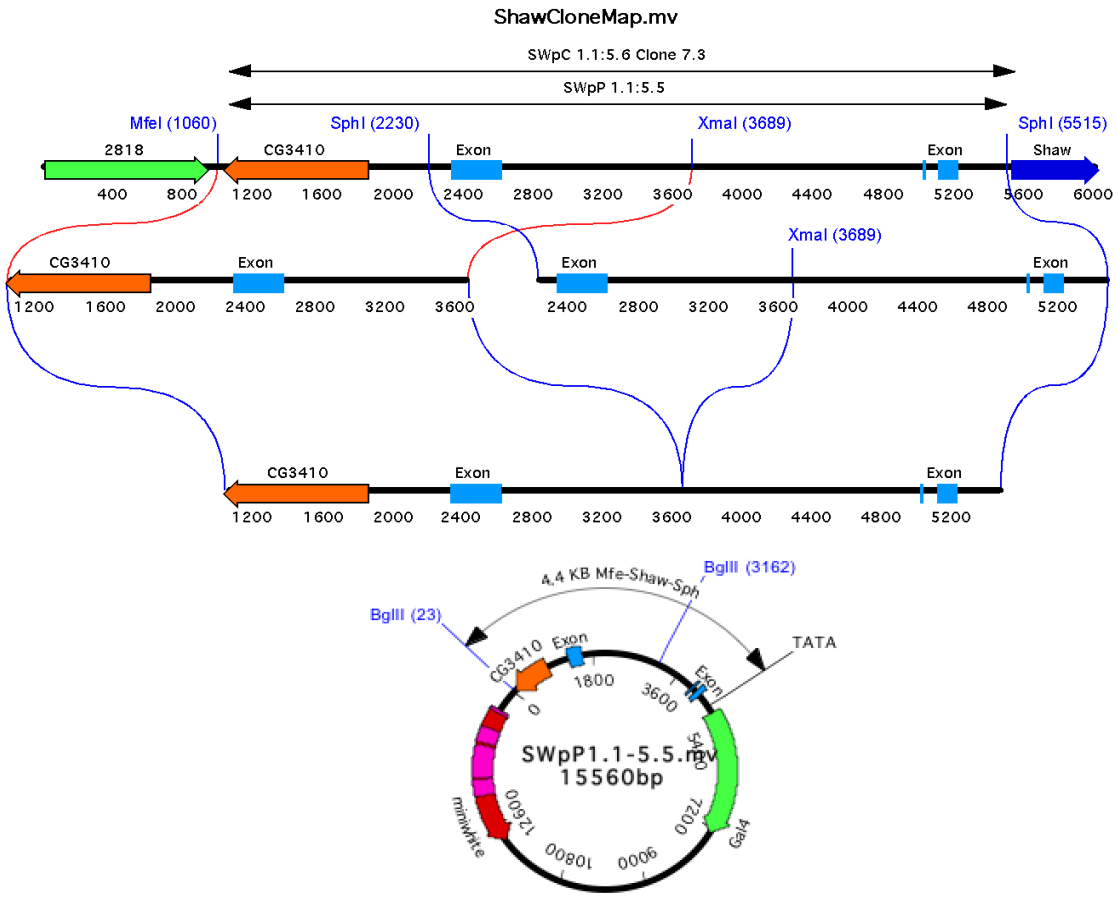


Figure 48: Brown sub-cloning scheme

A Mfe-Xma digest of 1.24 clone and a Sph I fragment of 1.24 are purified. Then the Sph I fragment is re-cut Xma and ligated to the Mfe-Xma fragment. This produces a Mfe-Sph fragment that can be sub-cloned to produce brown and olive.

Blue

This was generated from the excised fragment from a Sma I-BamH I digest of 1.24. The resulting gutted clone was end-filled, ligated and the PCR was run as the start of the "red" family of constructs (not reported here). The Sma-Bam fragment was ligated into an Nae-Bam digest of pHPelican and transformed.

Pink

This is simply a Bgl II-BamH I fragment (from the production of Aqua) that is ligated into the BamH I digested pHPelican.

Olive

Olive is the EcoR I fragment from dark orange ligated into the EcoR I site of pHPelican that has had the Bgl II- BamH I region removed from the multiple cloning site.

Reds

6 more constructs were produced to scan through the upper control region, but were not transformed in time. This will be continued later.

pPTGAL

All of the above constructs were produced in the pPTGAL vector, but were abandoned after the failure of the Kv construct lines.

REAL-TIME PCR

Total RNA from sex and age matched flies was extracted using either: the single-step RNA isolation protocol [internal ref], an in-house modification of this protocol [internal ref], or an RNeasy Kit (Part number 74104, Qiagen, Valencia, CA). The RNA

was treated with RNase-free DNase I (Ambion Inc. Austin, TX) to remove all DNA contamination. RNA was quantified using the RiboGreen® RNA Quantitation Kit (Part number R11490, Molecular Probes, Inc., Eugene OR) according to manufacturer instructions or with a Nanodrop Spectrophotometer.

First-strand cDNA was synthesized from typically 100 ng of total RNA, primed with 20 μM each of gene-specific primers for the gene being tested with Superscript II Reverse Transcriptase (Life Technologies, Gaithersburg, MD). The cDNA was amplified by Real-Time PCR, in an ABI Prism 7700 Sequence Detection System (Applied Biosystems, Foster City, CA), in the presence of gene specific dual-labeled single-stranded probes. PCR was performed using the TaqMan probes and the TaqMan Universal PCR Master mix (Applied Biosystems) or PlatinumTaq (Invitrogen, Austin, TX) polymerase and standard reagents. Each PCR was performed in triplicate and the yields thereof expressed as an average. mRNA abundance was quantified using the standard curve method. Significance was calculated using the Student's t-Test.

The Real-Time Primers used:

cyclophilin
Upper: 5' – ACCAACCACAACGGCACTG – 3'
Lower: 5' – TGCTTCAGCTCGAAGTTCTCATC – 3'
Probe: 5' – (FAM)–CGGCAAGTCCATCTACGGCAACAAGTT–(TAMRA) – 3'

shal
Upper: 5' – CCGTGTCTTCCGCATATTCA – 3'
Lower: 5' – ATGACGGTGGCAAAGATGATAA – 3'
Probe: 5' – (FAM)–TTCGGATCCTCGGCTA–(TAMRA) – 3'

shaw
Upper: 5' – GTCCTGGGCATCGTGATCTT – 3'
Lower: 5' – TAGCCGACGGTGGTCATTGT – 3'
Probe: 5' – (FAM)–CGCGGAGCGCAATCCAGCC–(TAMRA) – 3'

Shaw2 (CG4450)
Upper: 5' – GATACGCAAAACACCTTAGCCATATTA – 3'
Lower: 5' – AAGCTCGCCGTTGCTCACGT – 3'

Probe: 5’ – (FAM)–AACCAAATAGACGAGCTATCTGCTCCTCCG–(TAMRA) – 3’

Chapter 8: Flies

FLY STOCKS

Flies were raised on standard cornmeal/molasses/agar medium at 18˚C. Those meant for transformation and general purposes were maintained in the general population. Flies intended for *in situ* staining, Real Time PCR and immunohistochemical assays were kept on a 12 hour light/12 hour dark cycle with light starting at 9 am. When flies first started to eclose out of a food bottle, all the flies were cleared and new flies were then allowed to eclose over a 2 day period. They were then transferred to a fresh food bottle, and studied between 5 and 7 days later. For all experiments, unless otherwise noted, female flies were used.

Wild-type stocks were Canton-S and Oregon-R. Stocks used for transformation were w^{1118} for injections and the two stocks used for subsequent crosses were:

FM6/+; +; + (first chromosome balancers), and w-; cyo/sco; TM6/MKRS (second and third chromosome balancers).

Transformations

This protocol describes a procedure to transform Drosophila embryos with a reporter construct that will express the white (w+) gene as a marker of the successful incorporation of the plasmid into genomic DNA and the ß-galactosidase (lacZ) reporter gene driven by the transcriptional regulatory region cloned from genomic DNA.

DNA is prepared using a chemical-lysis midi-prep and then cleaned on a Qiagen column. This is quantified and then a preparation of 6 μg construct DNA is combined with 1 μg of Delta2-3 pTurbo helper plasmid (a transposase source) in 20 μl total of

injection buffer (0.1 M Sodium Phosphate Buffer pH 6.8, 5 mM KCl).

Embryos less than one hour in age were collected from egg plates using an embryo wash (in 1 L water, 6 g NaCl, 4 ml 10% Triton-X); de-chorionated with a 1.5 minute wash in 50% bleach. This is a time-critical step; any longer and survival rates will drop.

After affixing to a coverslip with a solution of hexane and 3M 445 tape glue, they were covered with 200 wt halocarbon oil (Halocarbon, Riveredge, NJ) that had been infused with oxygen for 3 hours. The embryos were injected with 5-10 pico-liters of injection solution at the posterior end of the syncitium (proximal to germ cells). Following the injection of a line of embryos on a coverslip, the glass was placed on an egg-laying plate and kept level. At 18°C, larvae emerged after 24-36 hours. These were collected and placed in a petri dish with cornmeal food base and a large dollop (~12 mls) of yeast paste. Pupae were collected after 5 days and placed in individual vials. After eclosion, the resulting flies are crossed to w^{1118} balancer stocks to isolate the insertion site of the transgene to a particular chromosome. This is beneficial to future crosses. Transformants are identified by the w+ phenotype. The eye color can range from deep red to caramel. Once transformant stocks are expanded, embryos are stained for ß-galactosidase activity, or crossed to UAS-reporter fly lines and visualized on the confocal microscope.

Egg Plates
Autoclave 960 ml H_2O and 25 g agar. A lower percentage results in agar contaminating the embryos washed off plates; a higher percentage results in fewer eggs laid.
Add 100 ml grape juice concentrate; 10 ml ethanol; 10 ml glacial acetic acid.

X-GAL STAINING (ß-GALACTOSIDASE ASSAY)

Embryos, along with larval and adult brains, were stained for ß-galactosidase activity with a modified procedure based on that published by Klambt (1991). Embryos were dechorionated with 50% bleach in water for 2.5 minutes and then fixed for 30 minutes in a 1.5 ml equal mixture of 3.7% formaldehyde in PBS and formaldehyde-saturated heptane. Alternatively, 5% paraformaldehyde in PBS (pH 7.2) was used. The fixative and heptane were removed and material was washed eight times, 15 minutes each in PBS with 0.3% Triton-X.

Staining was several hours to overnight at 37° C in Fe/Na Phosphate X-Gal solution: 10 mM NaPO4 pH 7.2; 150 mM NaCl; 1 mM MgCl2; 3 mM K4(FeII(CN)6); 3 mM K3(FeIII(CN)6); 0.3% Triton-X and 0.3% X-Gal (a range of 0.27%-0.33% was used). X-gal stock is 8% in DMSO. Both the salt and the X-gal stock solutions are heated to 65°C and then mixed together. Then the mixture is cooled to 37°C. After staining, one or two washes with PBS and 0.3% Triton-X will stop the reaction.

IN SITU HYBRIDIZATION

This protocol describes a procedure for the localization of mRNA with whole mount Drosophila embryos. It was also adapted for use with larval and adult brains. It is a modification of the Vectastain ABC labeling kit (VectorLabs, Burlingame, Ca). A digoxigenin labeled probe is prepared and then hybridized to whole mount specimens. A secondary antibody against digoxigenin conjugated to alkaline phosphatase is hybridized and a chemical reaction is used to show color for localization.

Preparation of the probe

Total RNA from 100 Canton-S flies was prepared, then an RT-PCR reaction performed using Superscript 2 and PlatinumTaq, with random hexamers and the primer pairs:

Shaw Upper 1
Upper: 5' – CTGATCAACATGGACTCGGAA– 3'
Shaw Ex1
Lower: 5' – AAACCGTCCGAAGAGGAATT – 3'

Shaw2 U6646
Upper: 5' – TAACTGCTGGCAGCGTATAAAA– 3'
Shaw2 L7408
Lower: 5' – CATAAAGTCGCCGGTCAATATT– 3'

Shal Ex2U
Upper: 5' – CCGGACGGGCAAGCTGCACTACCC– 3'
Shal EM0
Lower: 5' – TACTTCGACTCTTCGCCGCCCCCGATCG– 3'

These produced ~500 bp probes that were then twice gel-purified. Starting with 400 ng in 15 μl the probes were labeled with random primers and alkali-labile dUTP-Digoxigenin conjugate for 19 hours by adding 2 μl of random hexanucleotide primers; 2 μl of dNTP labeling mix (Roche [details]) and 1 μl Klenow enzyme. They were ethanol precipitated; brought up in 40 μl water and quantified. A final concentration of ~24 ng/μl (~1000 ng total) was generated for each probe. Of this, 45 μl of probe diluted in 155 μl of hybridization buffer (50% deionized formamide, 5X SSC, 100 μg/ml ssDNA, 200 μg.ml tRNA, and 0.1% Tween-20) was used in the *in situ*.

Preparation of whole mount material

Late stage (18-21 hour) embryos were collected from egg plates, rinsed with water and dechorionated for 2.5 minutes in 50% bleach and rinsed again with PBS, 0.3% Triton-X. These were transferred to a 1.5 ml microfuge tube containing 50% heptane and 50% fixative (3.7% formaldehyde, 50 mM EGTA in PBS (Phosphate Buffer Saline, 10x: 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO4, 2 mM KH_2PO_4) and fixed for 20 minutes on a rocking platform. The lower formaldehyde phase is removed and 1 ml of methanol is added to clear the embryos and remove the vitelline membrane. Shake vigorously for 15 seconds (we vortexed embryos for 30 seconds also) and let stand for 1 minute. Then remove the upper heptane layer and add more methanol. Repeat the methanol washes four times. Then wash four times in ethanol. Embryos are stored at -70° C to reduce background. They can be collected over time to accumulate enough material for the assay. They are then rehydrated in stages: 1 minute each step: 25% PBS-T (1X PBS, 0.1% Tween-20)/MeOH; 50% mix; 75-25% Mix.

We fixed the embryos again at this point; washed in PBS-T and incubated at 37° C for ~3 minutes with non-predigested proteinase-K (this is after running samples at different time points to evaluate the efficacy of the Proteinase-K). Stop the Proteinase-K activity with 2mg/ml glycine in PBS-T. Too long and the embryos disintegrate. Post-fix in 5% formaldehyde and 0.2% gluteraldehyde for 20 minutes. Then five 2 minute washes in PBS-T. Then remove 50% of PBS-T and add 50% hybe solution (50% deionized formamide, 5XSSC, 100 μg/μl ssDNA, 200 μg/μl tRNA, 0.1% Tween-20) to a 1:1 mixture. Then remove this and add 100% hybridization solution. Pre-hybridize 48° C for 2 hours.

Hybridization

Add 1 μg/ml heat denatured digoxygenin labeled probe (boil 5 minutes and then quick chill on ice). Hybridize at 48° C for 24-36 hours. Wash embryos in staged reduction of Hybridization solution (25% PBS-T: 75% Hyb. solution; 50% each; then 75% PSB-T: 25% Hyb. solution). Perform two washes of 20 minutes each with PBS-T. Alkaline Phosphatase-conjugated IgG Fraction Monoclonal Mouse Anti-Digoxin (Code# 200-52-156, JacksonResearch, West Grove, PA) is added at 1:2000 dilution of stock (600 μg/ml). Incubate for one hour and then wash four times in 20 minutes each in PBS-T.

Chemical Reaction

The native alkaline phosphatase activity must be blocked. Wash twice in 100 mM NaCl, 50 mM $MgCl_2$, 1 mM Levamisol (Sigma, L9756-5G), 0.1% Tween-20, for two minutes each. Add to the second 1 ml wash: 4.5 μl NBT and 3 μl X-Phosphate (BCIP, Roche, 10742020). Rock 5 minutes to several hours and watch as color develops. Stop reaction by washing at least six times in PBS-T. Wash once in 40% glycerol, then twice in 80% glycerol.

IMMUNOHISTOCHEMICAL STAINING

Preparation of whole mount material

Embryos were dechorionated with 50% bleach and then added to 700 μl heptane. An equal volume of: 3.7% formaldehyde in PEM Buffer (0.1 M PIPES, 1 mM $MgCl_2$ 1 mM EGTA, adjust pH 6.9 with KOH) was added and fixed for 20 minutes while rocking. Material was then washed 8 times, five minutes per wash, in PBTA (1x PBS, 1% BSA, 0.5% Triton-X, 0.02% Sodium Azide). In the case of embryos, the vitelline envelope

was cleared using the methanol wash from the *in situ* protocol. Pre-incubation is done for 30 minutes at room temperature in 1 ml PBT (1x PBS, 1% BSA, 0.05% Triton-X) with 5% Normal Donkey Serum (Code# 017-000-001, Jackson ImmunoResearch, West Grove, PA).

A similar preparation was used for larval brains and adult brains after dissection. For larvae, brains were dissected and cleaned of imaginal disks and trachea prior to fixing. For adult brains, whole flies were anesthetized with carbon dioxide gas and then immersed in ethanol for 30 seconds prior to dissection in PBS to remove the waxy cuticular coating. No heptane was used in the fixation, just the 3.7% formaldehyde in PEM buffer. Nor was a methanol clearing step used.

Hybridization

Two primary antibodies were used— a 1:300 dilution of goat anti-HRP (2.4 mg/ml) and a 1:1500 dilution of rabbit anti-ßgalactosidase (10 mg/ml stock solution) in the 1m PBS-T/Donkey Serum. Incubated overnight at 4° C. The primary solution was removed and stored at 4° C for usage later as preabsorbed. The samples were washed 8 times with PBTA (three rinses, and then five 15 minute washes). After the final wash, as much of the fluid as possible was removed and just enough secondary antibody solution (TRITC-Donkey anti-Goat [1:50 dilution of 1.5 mg/ml stock]; FITC-Donkey anti-Rabbit [1:100 dilution of 1.5 mg/ml stock]) was added to cover the samples. These were left at room temperature for 3 hours and then washed in PBTA 5 X 20 minutes.

Goat anti-HRP (123-905-021)

Rabbit anti-ßgalatosidase (Chemicon, AB1211-5MG)

Donkey anti-Goat TRITC conjugate (Jackson Immunoresearch 705-025-003)

Donkey anti-Rabbit FITC conjugate (Jackson Immunoresearch 711-095-152)

Chapter 9: Software

With the release of whole genome sequences, there has been an explosion of the number of tools available for analysis of these sequences— either *in toto* or with specific patterns in mind. These programs fall into three general categories: annotation tools such as national databases that allow for rudimentary searches and submission of descriptive infomation; pattern analysis tools that search the sequences for matches against specific external sequences of known importance; and tools that look for internal patterning of sequences (information content) that may have been preserved through time due to some functional constraint.

DATABASES

Eventually, all sequences were housed on local machines for analysis, however these were regularly reconciled with updated databases available on the Internet.

Flybase

This is a clearing house of information regarding Drosophila that is primarily housed at the University of California at Berkeley, but is mirrored at Harvard University, Indiana University, and in Europe at Cambridge University.

http://flybase.net/

The specific entries for the genes in this study are:

Kv1: http://www.flybase.org/.bin/fbidq.html?FBgn0003380
Kv2: http://www.flybase.org/.bin/fbidq.html?FBgn0003383
Kv4: http://www.flybase.org/.bin/fbidq.html?FBgn0005564
Kv3: http://www.flybase.org/.bin/fbidq.html?FBgn0003386
Kv3.2: http://www.flybase.org/.bin/fbidq.html?FBgn0032113

Transfac

Transfac is a database and search engine that utilizes curated transcription factor binding site data sets in discrete and matrix form. Raw sequence is matched against these databases with either the basic blast (Patch: the PAttern maTCH section) or a log-odds scoring matrix algorithm (Match: the MAtrix maTCH section). Transcription factor binding sequences are assigned a 'core' region of 5 base pairs that can be matched at a different stringency than the entirety.

This is a commercial project that is also available publicly as a subset of the total number of entries available to subscribing customers. With a subscription, the database and search engine are run locally. This allowed for the entry of additional matrices into the database. The Biobase (7.4 version) contained 38 insect entries. This was expanded to 62 locally using published data.

The public and commercial site is: http://www.gene-regulation.com/

Several other genome search algorithms now link sequence comparisons to the public version of this database (for example rVista).

On-line data bases.

There is no shortage of available databases on the internet. At this point, the list is expanding to the extent that any summation here would be futile. The reader is encouraged to use google.

SEQUENCE ANALYSIS

ClustalW

The source code for clustalW was obtained and modified and recompiled to accommodate the expanded symbols list used in the production of pseudosequences from transfac runs. The matrices were re-written with upper and lower case letters assigned to transcription factor binding site search results. A new identity matrix was created and linked to the runtime library.

Otherwise, the unmolested version of clustal 1.83 was used.

MEME

MEME or, Multiple Em for Motif Elicitation, was developed at the UC San Diego Engineering and Computer Science Department and is maintained jointly by UCSD, The University of Queensland, Australia, and National Center for Biological Resources (NCBR); and is available for download and installation (http://meme.sdsc.edu/meme/website/meme-download.html). MEME and Meta-MEME were successfully compiled and run on both a Linux platform and an Apple G4 locally.

The critical step for increasing the specificity of MEME is the training background files. The Release 3 annotations of the Drosophila genome are available as GFF files in raw text format. This made it possible to extract annotated coding (CDS), UTR, and repetitive elements- leaving predominantly non-coding sequence using standard bioperl scripts. These are then used as input to the Meta-MEME component, get-markov, which generates a third order markov model based on the fasta format input file.

fasta-get-markov < f.fasta > f.bg

MEME is then run with the argument:

--bg-file f.bg

This switch sets MEME to use the tuple frequencies of the bg file in converting the emission probabilities in the model to log-odds.

McPromoter

McPromoter is essentially Meta-MEME, that utilizes a Markov model generated from known promoters and data from Gibbs sampling of expected promoter sites based on their proximity to known transcription start sites. It runs as a distributed Linux executable on the Stuttgart Neural Network Simulator (http://www-ra.informatik.uni-tuebingen.de/SNNS/). Access was gained to the Texas Advanced Computing Center (http://www.tacc.utexas.edu/) and the code and background files installed for the Release 3 of the Drosophila Genome. The TACC Cray-Dell PowerEdge Xeon Cluster contains 768 3.06GHz and 256 3.2GHz Xeon processors within 512 Dell dual-processor PowerEdge 1750 compute nodes, 13 Dell dual-processor PowerEdge 2650 compute-I/O server-nodes and 2 Dell dual-processor PowerEdge 2650 login/management nodes. It is capable of 6.3 TFLOPS.

A 6 Kb sequence that includes the first coding exon of Kv3, the CG3410 transcript, and the last coding exon of the CG2818 gene from chromosome 2L was entered into the neural network engine and set for the highest sensitivity of 65% with a threshold of 0.8. Successive runs with thresholds of 0.9 and 0.95 were also run, with all three yielding the same promoter at the 5' UTR Exon of Kv3. McPromoter failed to find the minimal TATA-based promoter at Exon 1.

Uwe Ohler has made his HMM promoter analysis tool available via web interface at: http://genes.mit.edu/promoterMMII.html.

Macvector

Sequence annotation tool. Simple searches and graphical output of annotated sequences. Available commercially at an exhorbitant price.

Genomic Alignements

A resource for aligning whole genomes is available online at:

http://pipeline.lbl.gov/cgi-bin/gateway2

This in turn, is tied to a similar clearing house of genomic annotations at the University of California at Santa Cruz.

References

ADAMS, M. D. et al. The Genome Sequence of Drosophila Melanogaster.: The genome sequence of Drosophila melanogaster. Science, v. 287, n. 5461, p. 2185-2195, 2000.

ARNOSTI, D. N. Analysis and Function of Transcriptional Regulatory Elements: Insights From Drosophila.: Analysis and function of transcriptional regulatory elements: insights from Drosophila. Annu Rev Entomol, v. 48, n. p. 579-602, 2003.

BAILEY, T. L.; ELKAN, C. The Value of Prior Knowledge in Discovering Motifs With Meme.: The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol, v. 3, n. p. 21-29, 1995.

BARO, D. J. et al. Shab Gene Expression in Identified Neurons of the Pyloric Network in the Lobster Stomatogastric Ganglion: Shab gene expression in identified neurons of the pyloric network in the lobster stomatogastric ganglion. Receptors Channels, v. 2, n. 3, p. 193-205, 1994.

BARO, D. J.; COLE, C. L.; HARRIS-WARRICK, R. M. Rt-Pcr Analysis of Shaker, Shab, Shaw, and Shal Gene Expression in Single Neurons and Glial Cells.: RT-PCR analysis of shaker, shab, shaw, and shal gene expression in single neurons and glial cells. Receptors Channels, v. 4, n. 3, p. 149-159, 1996.

BATE, M.; MARTINEZ, A., ALFONSO The Development of Drosophila Melanogaster: The Development of Drosophila melanogaster. Plainview, N.Y: Cold Spring Harbor Laboratory Press, 1993

BEATY, B. J.; MARQUARDT, W. C. The Biology of Disease Vectors: The biology of disease vectors. Niwot, Colo: University Press of Colorado, 1996

BERMAN, B. P. et al. Exploiting Transcription Factor Binding Site Clustering to Identify Cis-Regulatory Modules Involved in Pattern Formation in the Drosophila Genome.: Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A, v. 99, n. 2, p. 757-762, 2002.

BETRAN, E.; LONG, M. Expansion of Genome Coding Regions By Acquisition of New Genes.: Expansion of genome coding regions by acquisition of new genes. Genetica, v. 115, n. 1, p. 65-80, 2002.

BIESSMANN, H. et al. Euchromatic and Heterochromatic Domains At Drosophila Telomeres.: Euchromatic and heterochromatic domains at Drosophila telomeres. Biochem Cell Biol, v. 83, n. 4, p. 477-485, 2005.

BLACKWOOD, E. M.; KADONAGA, J. T. Going the Distance: a Current View of Enhancer Action.: Going the distance: a current view of enhancer action. Science, v. 281, n. 5373, p. 60-63, 1998.

BOHM, R. A. et al. Transcriptional Control of Ca(2+)-Activated K(+) Channel Expression: Identification of a Second, Evolutionarily Conserved, Neuronal Promoter: Transcriptional control of Ca(2+)-activated K(+) channel expression: identification of a second, evolutionarily conserved, neuronal promoter. J Exp Biol, v. 203 Pt 4, n. p. 693-704, 2000.

BRAY, N.; DUBCHAK, I.; PACHTER, L. Avid: a Global Alignment Program.: AVID: A global alignment program. Genome Res, v. 13, n. 1, p. 97-102, 2003.

BRENNER, R.; ATKINSON, N. Developmental- and Eye-Specific Transcriptional Control Elements in an Intronic Region of a Ca(2+)-Activated K+ Channel Gene: Developmental- and eye-specific transcriptional control elements in an intronic region of a Ca(2+)-activated K+ channel gene. Dev Biol, v. 177, n. 2, p. 536-543, 1996.

BRENNER, R. et al. Tissue-Specific Expression of a Ca(2+)-Activated K+ Channel is Controlled By Multiple Upstream Regulatory Elements: Tissue-specific expression of a Ca(2+)-activated K+ channel is controlled by multiple upstream regulatory elements. J Neurosci, v. 16, n. 5, p. 1827-1835, 1996.

BUTLER, A.; WEI, A.; SALKOFF, L. Shal, Shab, and Shaw: Three Genes Encoding Potassium Channels in Drosophila: Shal, Shab, and Shaw: three genes encoding potassium channels in Drosophila. Nucleic Acids Res, v. 18, n. 8, p. 2173-2174, 1990.

BUTLER, J. E.; KADONAGA, J. T. Enhancer-Promoter Specificity Mediated By Dpe Or Tata Core Promoter Motifs.: Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. Genes Dev, v. 15, n. 19, p. 2515-2519, 2001.

——— The RNA Polymerase Ii Core Promoter: a Key Component in the Regulation of Gene Expression.: The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev, v. 16, n. 20, p. 2583-2592, 2002.

HOLMES, M. C.; TJIAN, R. Shaw Potassium Channel Genes in Drosophila.: Shaw potassium channel genes in Drosophila. Springer, 2000

CHANG, W. M. et al. Muscle-Specific Transcriptional Regulation of the

Slowpoke Ca(2+)-Activated K(+) Channel Gene: Muscle-specific transcriptional regulation of the slowpoke Ca(2+)-activated K(+) channel gene. J Biol Chem, v. 275, n. 6, p. 3991-3998, 2000.

COURONNE, O. et al. Strategies and Tools for Whole-Genome Alignments.: Strategies and tools for whole-genome alignments. Genome Res, v. 13, n. 1, p. 73-80, 2003.

DERMITZAKIS, E. T.; BERGMAN, C. M.; CLARK, A. G. Tracing the Evolutionary History of Drosophila Regulatory Regions With Models That Identify Transcription Factor Binding Sites: Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Mol Biol Evol, v. 20, n. 5, p. 703-714, 2003a.

DERMITZAKIS, E. T. et al. Evolutionary Discrimination of Mammalian Conserved Non-Genic Sequences (Cngs): Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science, v. 302, n. 5647, p. 1033-1035, 2003b.

ELEFANT, F. et al. Patterns of Histone Acetylation Suggest Dual Pathways for Gene Activation By a Bifunctional Locus Control Region.: Patterns of histone acetylation suggest dual pathways for gene activation by a bifunctional locus control region. EMBO J, v. 19, n. 24, p. 6814-6822, 2000.

GERASIMOVA, T. I.; CORCES, V. G. Chromatin Insulators and Boundaries: Effects on Transcription and Nuclear Organization.: Chromatin insulators and boundaries: effects on transcription and nuclear organization. Annu Rev Genet, v. 35, n. p. 193-208, 2001.

GHEZZI, A. et al. Slo K(+) Channel Gene Regulation Mediates Rapid Drug

Tolerance.: slo K(+) channel gene regulation mediates rapid drug tolerance. Proc Natl Acad Sci U S A, v. 101, n. 49, p. 17276-17281, 2004.

GINDHART, J. G. J.; KING, A. N.; KAUFMAN, T. C. Characterization of the Cis-Regulatory Region of the Drosophila Homeotic Gene Sex Combs Reduced.: Characterization of the cis-regulatory region of the Drosophila homeotic gene Sex combs reduced. Genetics, v. 139, n. 2, p. 781-795, 1995.

GONZALEZ, J. et al. Molecular Organization of the Drosophila Melanogaster Adh Chromosomal Region in D. Repleta and D. Buzzatii, Two Distantly Related Species of the Drosophila Subgenus.: Molecular organization of the Drosophila melanogaster Adh chromosomal region in D. repleta and D. buzzatii, two distantly related species of the Drosophila subgenus. Chromosome Res, v. 8, n. 5, p. 375-385, 2000.

GREGORY, P. D.; HORZ, W. Chromatin and Transcription--How Transcription Factors Battle With a Repressive Chromatin Environment.: Chromatin and transcription--how transcription factors battle with a repressive chromatin environment. Eur J Biochem, v. 251, n. 1-2, p. 9-18, 1998.

GREGORY, S. G. et al. A Physical Map of the Mouse Genome.: A physical map of the mouse genome. Nature, v. 418, n. 6899, p. 743-750, 2002.

GRUNDY, W. N. et al. Hidden Markov Model Analysis of Motifs in Steroid Dehydrogenases and Their Homologs.: Hidden Markov model analysis of motifs in steroid dehydrogenases and their homologs. Biochem Biophys Res Commun, v. 231, n. 3, p. 760-766, 1997.

GU, Z. et al. Extent of Gene Duplication in the Genomes of Drosophila, Nematode, and Yeast.: Extent of gene duplication in the genomes of Drosophila,

nematode, and yeast. Mol Biol Evol, v. 19, n. 3, p. 256-262, 2002.

HO, Y. et al. A Defined Locus Control Region Determinant Links Chromatin Domain Acetylation With Long-Range Gene Activation.: A defined locus control region determinant links chromatin domain acetylation with long-range gene activation. Mol Cell, v. 9, n. 2, p. 291-302, 2002.

HODGE, J. J. et al. Shaw Potassium Channel Genes in Drosophila.: Shaw potassium channel genes in Drosophila. Curr Opin Genet Dev, v. 6(6), n. 6, p. 235-254, 2005.

HODGKIN, J. A View of Mount Drosophila: A view of mount Drosophila. Nature, v. 404, n. 6777, p. 442-443, 2000.

HOLMES, M. C.; TJIAN, R. Promoter-Selective Properties of the Tbp-Related Factor Trf1.: Promoter-selective properties of the TBP-related factor TRF1. Science, v. 288, n. 5467, p. 867-870, 2000.

KENT, W. J. Blat--the Blast-Like Alignment Tool.: BLAT--the BLAST-like alignment tool. Genome Res, v. 12, n. 4, p. 656-664, 2002.

KRAFT, P.; HORVATH, S. The Genetics of Gene Expression and Gene Mapping.: The genetics of gene expression and gene mapping. Trends Biotechnol, v. 21, n. 9, p. 377-378, 2003.

LEE, M. P. et al. Atg Deserts Define a Novel Core Promoter Subclass.: ATG deserts define a novel core promoter subclass. Genome Res, v. n. p. 2005.

LEMON, B.; TJIAN, R. Orchestrated Response: a Symphony of Transcription Factors for Gene Control.: Orchestrated response: a symphony of transcription factors for

gene control. Genes Dev, v. 14, n. 20, p. 2551-2569, 2000.

LEWIS, E. B. The Bithorax Complex: the First Fifty Years.: The bithorax complex: the first fifty years. Int J Dev Biol, v. 42, n. 3, p. 403-415, 1998.

LOOTS, G. G. et al. Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 By Cross-Species Sequence Comparisons.: Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science, v. 288, n. 5463, p. 136-140, 2000.

LORENZINI, E. C.; SCHEFFLER, I. E. Co-Operation of the 5' and 3' Untranslated Regions of Ornithine Decarboxylase Mrna and Inhibitory Role of Its 3' Untranslated Region in Regulating the Translational Efficiency of Hybrid RNA Species Via Cellular Factor.: Co-operation of the 5' and 3' untranslated regions of ornithine decarboxylase mRNA and inhibitory role of its 3' untranslated region in regulating the translational efficiency of hybrid RNA species via cellular factor. Biochem J, v. 326, n. Pt 2, p. 361-367, 1997.

MCCORMACK, T. J. Comparison of K+-Channel Genes Within the Genomes of Anopheles Gambiae and Drosophila Melanogaster.: Comparison of K+-channel genes within the genomes of Anopheles gambiae and Drosophila melanogaster. Genome Biol, v. 4, n. 9, p. R58, 2003.

MERIKA, M.; THANOS, D. Enhanceosomes.: Enhanceosomes. Curr Opin Genet Dev, v. 11, n. 2, p. 205-208, 2001.

MILLER, W. Comparison of Genomic DNA Sequences: Solved and Unsolved Problems: Comparison of genomic DNA sequences: solved and unsolved problems. Bioinformatics, v. 17, n. 5, p. 391-397, 2001.

MINSHALL, N.; STANDART, N. The Active Form of Xp54 RNA Helicase in Translational Repression is an RNA-Mediated Oligomer.: The active form of Xp54 RNA helicase in translational repression is an RNA-mediated oligomer. Nucleic Acids Res, v. 32, n. 4, p. 1325-1334, 2004.

MISRA, S. et al. Annotation of the Drosophila Melanogaster Euchromatic Genome: a Systematic Review.: Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biol, v. 3, n. 12, p. RESEARCH0083, 2002.

MODREK, B.; LEE, C. J. Alternative Splicing in the Human, Mouse and Rat Genomes is Associated With an Increased Frequency of Exon Creation and/Or Loss.: Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet, v. 34, n. 2, p. 177-180, 2003.

NABAUER, M.; KAAB, S. Potassium Channel Down-Regulation in Heart Failure.: Potassium channel down-regulation in heart failure. Cardiovasc Res, v. 37, n. 2, p. 324-334, 1998.

OHLER, U. et al. Computational Analysis of Core Promoters in the Drosophila Genome.: Computational analysis of core promoters in the Drosophila genome. Genome Biol, v. 3, n. 12, p. RESEARCH0087, 2002.

OHLER, U. et al. Stochastic Segment Models of Eukaryotic Promoter Regions.: Stochastic segment models of eukaryotic promoter regions. Pac Symp Biocomput, v. n. p. 380-391, 2000.

PARO, R. Imprinting a Determined State Into the Chromatin of Drosophila.: Imprinting a determined state into the chromatin of Drosophila. Trends Genet, v. 6, n. 12, p. 416-421, 1990.

PAYNE, C. M.; MULLINS, L. J.; MULLINS, J. J. Manipulating Large Genomic Clones Via in Vivo Recombination in Bacteria.: Manipulating large genomic clones via in vivo recombination in bacteria. J Hum Hypertens, v. 13, n. 12, p. 845-848, 1999.

PFEUFER, A. et al. Common Variants in Myocardial Ion Channel Genes Modify the Qt Interval in the General Population: Results From the Kora Study.: Common variants in myocardial ion channel genes modify the QT interval in the general population: results from the KORA study. Circ Res, v. 96, n. 6, p. 693-701, 2005.

POGWIZD, S. M. et al. Arrhythmogenesis and Contractile Dysfunction in Heart Failure: Roles of Sodium-Calcium Exchange, Inward Rectifier Potassium Current, and Residual Beta-Adrenergic Responsiveness.: Arrhythmogenesis and contractile dysfunction in heart failure: Roles of sodium-calcium exchange, inward rectifier potassium current, and residual beta-adrenergic responsiveness. Circ Res, v. 88, n. 11, p. 1159-1167, 2001.

RUBIN, G. M. et al. Comparative Genomics of the Eukaryotes: Comparative genomics of the eukaryotes. Science, v. 287, n. 5461, p. 2204-2215, 2000.

SCHNEIDER, H. et al. Patterns of Shaker Family Gene Expression in Single Identified Neurons of the American Lobster, Homarus Americanus.: Patterns of shaker family gene expression in single identified neurons of the American lobster, Homarus americanus. Receptors Channels, v. 7, n. 1, p. 53-64, 2000.

SHABALINA, S. A. et al. Comparative Analysis of Orthologous Eukaryotic Mrnas: Potential Hidden Functional Signals.: Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucleic Acids Res, v. 32, n. 5, p. 1774-1782, 2004.

SIRONI, M. et al. Analysis of Intronic Conserved Elements Indicates That Functional Complexity Might Represent a Major Source of Negative Selection on Non-Coding Sequences.: Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. Hum Mol Genet, v. 14, n. 17, p. 2533-2546, 2005.

SMALE, S. T. Core Promoters: Active Contributors to Combinatorial Gene Regulation.: Core promoters: active contributors to combinatorial gene regulation. Genes Dev, v. 15, n. 19, p. 2503-2508, 2001.

SMALL, S.; BLAIR, A.; LEVINE, M. Regulation of Even-Skipped Stripe 2 in the Drosophila Embryo.: Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J, v. 11, n. 11, p. 4047-4057, 1992.

SMIT, A. F. The Origin of Interspersed Repeats in the Human Genome.: The origin of interspersed repeats in the human genome. Curr Opin Genet Dev, v. 6, n. 6, p. 743-748, 1996.

TAMKUN, J. W. et al. Brahma: a Regulator of Drosophila Homeotic Genes Structurally Related to the Yeast Transcriptional Activator Snf2/Swi2.: brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. Cell, v. 68, n. 3, p. 561-572, 1992.

TUPLER, R.; PERINI, G.; GREEN, M. R. Expressing the Human Genome.: Expressing the human genome. Nature, v. 409, n. 6822, p. 832-833, 2001.

VENTER, J. C. et al. The Sequence of the Human Genome.: The sequence of the human genome. Science, v. 291, n. 5507, p. 1304-1351, 2001.

WASSERMAN, W. W. et al. Human-Mouse Genome Comparisons to Locate

Regulatory Sites: Human-mouse genome comparisons to locate regulatory sites. Nat Genet, v. 26, n. 2, p. 225-228, 2000.

WIEHE, T.; GUIGO, R.; MILLER, W. Genome Sequence Comparisons: Hurdles in the Fast Lane to Functional Genomics.: Genome sequence comparisons: hurdles in the fast lane to functional genomics. Brief Bioinform, v. 1, n. 4, p. 381-388, 2000.

WINSLOW, R. L.; RICE, J.; JAFRI, S. Modeling the Cellular Basis of Altered Excitation-Contraction Coupling in Heart Failure.: Modeling the cellular basis of altered excitation-contraction coupling in heart failure. Prog Biophys Mol Biol, v. 69, n. 2-3, p. 497-514, 1998.

WONG, G. K. et al. Is "Junk" DNA Mostly Intron DNA?: Is "junk" DNA mostly intron DNA? Genome Res, v. 10, n. 11, p. 1672-1678, 2000.

WU, C. H. et al. Analysis of Core Promoter Sequences Located Downstream From the Tata Element in the Hsp70 Promoter From Drosophila Melanogaster.: Analysis of core promoter sequences located downstream from the TATA element in the hsp70 promoter from Drosophila melanogaster. Mol Cell Biol, v. 21, n. 5, p. 1593-1602, 2001.

ZHU, Y. Y. et al. Reverse Transcriptase Template Switching: a Smart Approach for Full-Length Cdna Library Construction.: Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques, v. 30, n. 4, p. 892-897, 2001.

Vita

Moon Draper arrived to the study of biology through a circuitous path fueled by intense curiosity and stubbornness. Raised in Kansas by kind folk, Moon left the midwest to attend several universities gaining degrees in Earth Sciences and, eventually, a Bachelor of Science in Biology at the University of Texas at Arlington. In the interim, Moon worked in a variety of professions and travelled. He started graduate school in 1996. While pursuing research for his doctorate at the University of Texas at Austin, Moon rediscovered his love of teaching.

Following the completion of this work, Moon will return to what he loves.

Permanent address: PO Box 180911 Austin, TX 78718-0911

This dissertation was typed by Moon Draper.