CHAPTER 3

ANALYSIS OF CG32062 EXPRESSION PATTERN

CG32062- Drosophila homolog of A2BP1 group of proteins
Expression pattern of CG32062
CHAPTER 3

ANALYSIS OF CG32062 EXPRESSION PATTERN

3.1. INTRODUCTION

3.1.1. Enhancer Trap Approach and EN403-GAL4

A rapid improvement in unraveling the developmental and regulatory role of number of genes was made possible by the inventions and application of genetic screens based on chemical, irradiation or P-element mutagenesis. Though chemical mutagenesis and irradiation methods create mutation at a higher frequency, characterizing the gene disrupted randomly is laborious. An improved method of P-element mutagenesis called the enhancer-trap is convenient as the histological system in this method reveals the probable expression pattern of the gene of interest (Pignoni et al., 1997). In Drosophila, this enhancer-trap is widely used to dissect the tissue-specific gene expression, wherein the P-element containing the lacZ reporter gene with a minimal promoter is mobilized into the genome using a transposase source. Since the P-element is sensitive to the regulatory elements at the region of insertion, reporter gene activity depicts the endogenous gene activity spatio-temporally (O’Kane and Gehring 1987; Bellen et al. 1990).

GAL4-enhancer trap method is a derivative of the enhancer-trap technique, in which GAL4 transcription activator is cloned under the control of the local regulatory region in the P-element construct (Brand and Perrimon 1993). This system is useful to drive the expression of any gene of interest placed downstream of the upstream activation sequence (UAS). The GAL4 system has now become a standard tool in Drosophila genetic analyses. The low efficiency of obtaining number of GAL4 lines has been overcome by the targeted transposition technique, where a precise exchange of GAL4 carrying P element takes place with a pre-existing P-element (Sepp and Auld, 1999).
Earlier in our lab, to identify novel markers that are potential targets of Homeotic gene control in Drosophila, enhancer trap approach was employed (Bajpai et al., 2004). From a collection of 800 such GAL4 enhancer-trap lines, EN403-GAL4, a line showing segmentally modulated pattern in its expression in the larval CNS was identified. In wing and haltere imaginal discs, EN403 is expressed in a subset of non-D/V cells of the pouch, on either side of the D/V boundary. Its expression in the wing pouch is more prominent in proximal than in distal cells. EN403 is also expressed in the peripodial cells of wing imaginal discs. EN403 is differentially expressed between wing and haltere discs. In wing discs, it is expressed in both anterior and posterior compartments, while in haltere discs, it is expressed only in the posterior compartment. EN403 is also expressed in the presumptive distal segments of leg and antennal discs and in the morphogenetic furrow of eye discs. By Polytenic in situ hybridization and sequence analysis, EN403 GAL4 was mapped to 67E2-E3, within the 50kb long 2nd intron of CG32062. The gene codes for a RNA-binding protein and its closest homologue in vertebrates, at the level of RNA binding motif, is human Ataxin 2 binding protein 1 (A2BP1). By comparative study of RNA in situ hybridization using a cDNA for CG32062 and EN403 pattern, it was inferred that EN403 is an enhancer trap of CG32062 (Bajpai et al., 2004). Using targeted P-conversion approach (Sepp and Auld, 1999), P element EN403-GAL4 was exchanged with a ptc-lacZ line to generate EN403-lacZ. EN403-lacZ shows an expression pattern identical to that of EN403-GAL4 in wing discs and in all other tissues.

3.1.2. RNA Binding Proteins

In eukaryotes, post-transcriptional regulation of gene expression is achieved by numerous RNA-binding proteins, which can be classified into several groups. RNA-binding proteins are important players in the regulation of gene expression (Nagai, 1996). RNA-binding proteins are involved in important developmental decisions made at the level of mRNA processing, stability, localization or translational control. RNA is never a lonely molecule! As soon as it is transcribed, number of ribonucleotides (RNPs) bind to nascent RNA and aid in a number of cellular activities. Similarly, specific RNA-protein interactions play crucial roles in gene regulation through transcriptional control, RNA
processing, transport and translational control. Three commonly found RNA-binding domains are the ribonucleoprotein (RNP) domain, the double stranded RNA binding domain (dsRBD) and the K homology (KH) domain. The three dimensional structure of small RNA-binding domains showed that this domain contains two short conserved sequence motifs called RNP1 (RNP octamer) and RNP2 (RNP hexamer) within a weakly conserved motif made up of about 70-90 amino acid residues. The RNP domain includes a four-stranded antiparallel beta-sheet flanked on one side by two alpha helices. The RNP1 and RNP2 motifs are located in the two middle beta strands and play a crucial role in RNA binding (Fig.3.1).

Fig.3.1. Crystal Structure of RNA Recognition Motif (RRM). Structure of the N-terminal RNA-recognition motif (RRM) of human U1A bound to RNA is shown. In RRM–RNA complexes, single-stranded bases are specifically recognized through the protein β-sheet and through two loops that connect the secondary structure elements.
The function of RNA-binding proteins is particularly important during embryonic development when complex spatial and temporal programs of gene expression are established (Colegrove-Otero et al., 2005). Indeed, RNA-binding proteins have been shown to be required for axis specification, germ plasm formation, sex determination or determination of blastomere identity (de Moor et al., 2005; Minakhina and Steward, 2005; Wilhelm and Smibert, 2005).

Alternative splicing, one of the central mechanisms that regulate eukaryotic gene expression, produces functionally distinct proteins from a single gene. This mechanism has a prime role in gene regulation both during development and differentiation. Many of the proteins with RNA binding domain are closely related to alternate splicing mechanism. Most striking example of this type is reflected in the *Drosophila* sex determination pathway involving Sex-lethal (Sxl) protein with RNA-Binding Domain (RBD) of RRM type. Sxl protein with two RBDs has been known not only to regulate alternative splicing but also translation of genes in sex determination pathway. Sxl recognizes the sequence context downstream of the uridine-rich sequence during alternative splicing and Sxl RNA binding requires both of its RBDs for specific activity (Sakashita and Sakamoto, 1996; Samuels et al., 1998). It has been found that Sxl RBDs mediate homodimeric interactions between two Sxl molecules, indicating the possibility of protein-protein interaction. Such an interaction between Sxl molecules is enhanced and facilitated by specific RNA molecules with Sxl-binding sites that bring about conformational changes.
To date, numerous RRM-type RNA-binding proteins containing multiple RBDs have been identified. Some have been shown to bind specific RNA sequences via a single RBD. Unlike Sxl protein, where both RBDs are important for its function, the human snRNP protein U1A, requires only its first RRM for binding to U1 snRNA. In this paradigm, each of its RBD, if taken individually, is found to function differently and display different RNA binding specificities (Lutz-Freyermuth et al., 1990; Tsai et al., 1991). This demonstrates that RNA binding specificity is not merely the sum of the binding specificities of each of the individual RBDs, as in some of the RRM-type RBDs, cooperative action of multiple intramolecular RBDs appear to be important to specify the RNA binding function.

Similarly in a number of functionally important neuronal genes, alternative splicing has been found to be a common mechanism adopted to generate tissue specific isoforms. Such a phenomenon was first observed in CA/CGRP gene encoding two peptides that differ in the terminal exon, one, calcitonin specific for thyroid cells and CGRP in specific neurons (Amara et al. 1982; Crenshaw et al. 1987). In Drosophila, ELAV is a classical example of a nuclear protein, expressed specifically in neurons, with two tandem RRM type domain, followed by a hinge domain and a terminal RRM, and implicated in splicing and auto-regulation (Birney et al. 1993; Yannoni and White 1997; Samson 1998; Lisbin et al. 2000). In vertebrates too, many genes are shown to be regulated by alternative splicing. For example, a neuron specific RNA-binding protein, Nova-1, regulates alternative splicing in neurons, wherein nova-1 null mice possess a specific splicing defect in the inhibitory glycine receptor α2 (GlyRα2) exon 3A(Jensen et al., 2000).

Fox-1 is a classical example of a protein, conserved in Caenorhabditis elegans and Zebrafish, which is homologous to human and mouse A2BP, that possesses RRM-type RNA binding protein, and found to function by engaging tissue specific isoforms. In zebrafish, fox-1 is expressed during muscle development, whereas a mouse homologous gene is expressed in brain as well as in heart and skeletal muscles. An intronic hexanucleotide UGCAUG has been shown to play a critical role in the regulation of tissue specific alternative splicing of pre-mRNAs in a wide range of tissues. Vertebrate
Fox-1 has been shown to bind to this element, in a highly sequence-specific manner, through its RNA recognition motif (RRM) (Jin et al., 2003; Nakahata and Kawamoto, 2005; Lunde et al., 2007). In mammals, there are at least two Fox-1-related genes, ataxin-2 binding protein 1 (A2BP1)/Fox-1 and Fxh/Rbm9, which encode an identical RRM. It has been demonstrated that both mouse Fxh and A2BP1 transcripts undergo tissue-specific alternative splicing, generating protein isoforms specific to brain and muscles. These tissue-specific isoforms are characterized for their abilities to regulate neural cell specific alternative splicing of a cassette exon, N30, in the non-muscle myosin heavy chain II-B pre-mRNA, which is regulated through an intronic distal downstream enhancer (IDDE). All Fxh and A2BP1 isoforms with the RRM have been demonstrated to be capable of binding to the IDDE in vitro through the UGCAUG elements. All Fxh isoforms and a brain isoform of A2BP1 show a predominant nuclear localization. Brain isoforms of both Fxh and A2BP1 are found to be more efficient to promote N30 splicing than the muscle specific isoforms. Skeletal muscles isoforms found to lack a part of the RRM, are found not only to be incapable of activating neural cell specific splicing but found to inhibit UGCAUG-dependent N30 splicing. These findings suggest that tissue-specific isoforms of Fxh and A2BP1 play an important role in determining tissue specificity of UGCAUG-mediated alternative splicing. Study on zebrafish homolog of Caenorhabditis elegans Fox-1 has shown that it could bind specifically to the pentanucleotide GCAUG. This pentanucleotide is almost identical to the hexanucleotide UGCAUG except for the first U. The zebrafish Fox-1 homolog, as well as the mouse Fox-1 homolog, is found to be capable of repressing the inclusion of an alternative cassette exon of the ATP synthase Flg pre-mRNA via binding to GCAUG, that is similar to muscle-specific exclusion of this exon.

In Drosophila, there are four major classes of RNA binding proteins, namely, RNA recognition motif (RRM) proteins, DEAD/DExH-box helicases, KH domain proteins, and double-stranded RNA binding domain (DSRBD) proteins. Of all, RRM type of proteins is the largest with 117 proteins (reviewed Paul Lasko, 2000). Many of them are found to possess dual functionality, like Bicoid, that possess both DNA and RNA binding function. Many of these RBD proteins have homology to other eukaryotic proteins. For
example CG6049 of *Drosophila* encodes a protein similar to human Tat stimulatory factor-1 (Tat-SF1), a cofactor of Tat (Li and Green, 1998). Tat, in cooperation with Tat-SF1 and another protein TAR, functions as a transcription elongation factor of HIV-1, by forming a stem-loop structure at the 5' end of the nascent viral transcript. In addition, in *Drosophila* genome there are relatively large number of other RNA binding proteins without specific domains, whose functions are yet to be predicted and analysed. For example, Apontic is a protein without any known RBD, but yet found experimentally to possess RNA binding function (Lie and Macdonald, 1999).

Differential expression of protein isoforms with conserved RNA-binding domains with different functions is yet another productive way of regulating the gene expression during development of an organism. Several isoforms of CPEB, PABP, Pumilio and Staufen proteins, that are differentially expressed, are found to perform different regulatory functions both in vertebrates as well as in invertebrates. In many cases, additional modification of these RNA-binding proteins in response to activation of certain signaling pathways is employed to regulate its function. For example, in Xenopus oocytes hnrNPa1 is phosphorylated by PKA to localize the protein in the cytoplasm, while, during meiotic maturation when PKA is downregulated, hnrNPa1 is relocated to the nucleus (Xie *et al.*, 2003). Thus, phosphorylation modifies the proteins with RBDs in localization as well as stabilization and thereby aid in its regulation of translation of other proteins.

### 3.1.3. Poly (Q)-mediated Neurodegenerative Diseases

Polyglutamine (PolyQ) diseases, a growing class of inherited neurodegenerative disease, includes these following representative diseases - Huntington's disease (HD), spinal and bulbar muscular atrophy, dentatorubral-pallidoluysian atrophy (DRPLA), spinocerebellar ataxia type 1 (SCA1), SCA2, spinocerebellar ataxia type 3/ Machado-Joseph disease, spinocerebellar ataxia type 6 and spinocerebellar ataxia type 7. The genes share no homology except for the CAG repeats encoding polyglutamine. A common feature in the pathology of the triplet repeat disease is the presence of either ubiquitinated intranuclear inclusions (NII) composed of aggregated proteins or cytoplasmic aggregates. The
expanded polyQ stretch confers toxic properties on the disease proteins through alteration of their conformation leading to pathogenic protein–protein interactions including oligomerization and/or aggregation.

In Spinocerebellar ataxia 3 / Machado Joseph Disease (SCA3/MJD) polyglutamine expansion results in altered, misfolded domain within this protein, more importantly involving intranuclear aggregates or nuclear inclusions (NII) (Chai et al., 1999). In contrast, study by Huynh et al., in 2000 on SCA2 patient brain found no NII, and it suggested that the formation of NII is not essential for the pathogenesis of HD or of SCA1.

There are different schools of thought and hypothesis on the development of poly(Q) diseases. In addition to number of proteins involved in the pathogenecity of the disease, possible role of RNA in disease development was also found as in Myotonic dystrophy (DM1), a multisystemic neurological disease characterized by progressive muscle weakness, ocular cataracts and cardiac arrhythmias. In DM1, as the CTG expansion is located in the 3′-untranslated region (UTR) of a protein kinase gene, the location of the mutation suggested that the repeat expansion may not be toxic at the protein level but at the level of RNA synthesis, processing or transport. Several hypotheses have been proposed to explain dominant type inheritance in DM1 by gain-of-function at the RNA level (reviewed by Broude and Cantor, 2003).

Comparison of the increased susceptibility of cytoplasmic and nuclear polyglutamine aggregates to autophagic degradation, reveals autophagy to be a critical component of the cellular clearance of toxic protein aggregates suggesting protein aggregates to be more toxic when directed to the nucleus (Iwata et al 2005).

To take the advantage of the power of Drosophila genetics to address the problem of PolyQ-mediated neurodegeneration, Drosophila models of this class of diseases have been generated. A number of Drosophila models are available for PolyQ diseases, which can be used to unravel mechanisms of degeneration and gradual neuronal loss. Towards
this end, it is possible to perform genetic screens to find out genes that modulate the neurodegeneration disease, characterize and validate them. For example, the human gene encoding Hsp70, a molecular chaperone that modulates protein folding, was introduced into *Drosophila* to address whether raising Hsp70 levels could combat the toxic effects of polyQ. It was found that co-expression of human Hsp70 ameliorated polyQ toxicity and suppressed degeneration.

Fly models have been used to understand the development of neurodegenerative disease in humans and thereby develop ways to prevent the same. In the *Drosophila* model of SCA3yMachado–Joseph disease, it was found while wildtype Q27 protein had no effect, Q78 protein induced late-onset, progressive degeneration, with the formation of nuclear inclusions (reviewed by Paulson et al., 2000). In another disease model of Spinocerebellar ataxia type 1 (SCA1), it was found that expression of normal-length human SCA1 protein in *Drosophila* was sufficient to develop the degenerative phenotypes similar to those caused by the expanded protein (Pedro et al., 2000). In another study, transgenic *Drosophila* expressing full-length ataxin-1 (30Q-wildtype human isoform) and ataxin-1 (82Q an expanded isoform) were generated and using UAS/GAL4 modulatory system the effect of SCA1 expression in retinal neurons of eyes were studied. The nuclear inclusions were found in a variety of cell types like photoreceptor cells, neurons of CNS wherever 30Q or 82Q ataxin-1 was expressed and it was found that the expression of ataxin-1 in fly neurons generated progressive degeneration as in patients with SCA1 and transgenic mice.

Suppressor-enhancer and modifier screen has been carried out in *Drosophila* to identify the genes and pathways involved in polyglutamine-induced degeneration of SCA1. A suppressor of SCA1 neurodegeneration, *glutathione-s-transferase* (*GST*), that probably mediate cellular detoxification and modifier of SCA1 neurodegeneration, *mushroom-body expressed* (*mub*) that encode protein containing RNA-binding KH domains that could stabilize specific messenger RNA were identified. Interestingly other enhancers identified in the screen were the genes like *pum, cpo, dyT521-B*, all of which encode for proteins containing RNA binding domains. This suggested that alteration of RNA
processing might occur during SCA1 pathogenesis. Additionally, a second group of enhancer proteins like Sin3A, Rpd3 and dCtBP that function as cofactors in transcriptional regulation were also identified (Fernandez-Funez et al., 2000). These modifiers suggest that abnormal interactions between ataxin-1 and the transcriptional machinery might be another pathway for pathogenesis. They also enforce the role of protein folding and protein clearance in the development of the disease. Not only does this discovery of genes involved in cellular detoxification, transcription regulation and RNA processing suggest additional pathogenic mechanisms, but also help in treating such polyglutamine and neurodegenerative diseases and other neurodegenerative disorders such as Alzheimer’s disease and Parkinson’s disease.

Using Drosophila models, it is also possible to identify inhibitors or proteins that combat poly(Q) diseases and to check the effect of external inhibitors. Many polyQ protein-interacting partners that are involved in disease pathogenesis, such as Huntingtin-associated protein 1, GAPDH and CREB binding protein (CBP) have been reported. Molecules with selective binding affinity to the expanded polyQ stretch and those that interfere with the pathogenic properties, like Polyglutamine Binding Peptide 1 and 2 (QBP1 and QBP2) have been tested for their inhibitory roles in Drosophila disease models (Nagai, et. al. 2003). The inhibitor peptide QBP1 was found to significantly suppress polyQ aggregation and polyQ induced neurodegeneration. Thus, the peptide inhibitor QBP1 appears to be a promising candidate with a potential therapeutic value against the currently untreatable polyQ diseases.

Identification of the signaling pathway with which these expanded polyglutamine peptides interact is another way of combating the disease. Results from a study that polyQ-containing domain of protein huntingtin (Htt) antagonizes the EGFR-mediated ERK signaling pathway in Drosophila glial cells and that polyQ functions upstream to this pathway, suggested that disruption of EGFR signaling and dysfunctioning the glial cells by polyQ expansion could probably be the potential routes of pathogenesis of HD and other polyQ diseases in humans (Lievens et al., 2005). Thus, Drosophila disease models reflect all the basic phenomenon of human polyQ diseases that involves number
of neural defects, its gradual degeneration and accumulation of abnormal protein aggregations.

### 3.1.4. Ataxin-2 and A2BP

Ataxin-2, the SCA2 gene product is a highly basic protein (mol.wt. 145 kDa) except for one acidic domain containing 50 amino acids, downstream of the polyQ tract. Added to several other functional motifs within this domain, the most common normal SCA2 alleles contain 22 or 23 CAG repeats with one or two CAA interruptions, whereas the repeat is expanded to 34 repeats without interruption in disease alleles. Ataxin-2 is highly conserved in evolution with orthologs in mouse, *Caenorhabditis elegans* and *Drosophila melanogaster*. Though disease development of SCA2 is restricted to the cerebellum and brainstem, the *sca2* transcript is found to be widely expressed in various other tissues like heart, placenta, liver, skeletal muscle, pancreas and brain. Immunocytochemistry of normal and SCA2 brains showed that ataxin-2 expression was localized to the cytoplasm of specific neuronal groups, but not restricted to Purkinje cells, which are the primary targets of SCA2 pathogenesis. Huynh *et al.* (2000) studied the disease pathogenesis in cell cultures, human and mouse SCA2 brains and found the expanded polyQ proteins to be cytoplasmic, but not nuclear aggregation. Mice expressing ataxin-2 with Q58 showed early onset of functional deficits, abnormalities and loss of Purkinje cells and its dendritic branches. Ataxin-2 does not contain any of the known nuclear localization signals and has an exclusively cytoplasmic localization in normal and SCA2 human brain.

In 2002, Satterfield used genetic and molecular approaches to investigate the function of a *Drosophila* homolog of the SCA2 gene (Datx2). Like human ataxin-2, Datx2 is found to be active throughout development in a variety of tissue types and localized to the cytoplasm. Mutations that reduced Datx2 activity or transgenic overexpression of Datx2 resulted in female sterility, aberrant sensory bristle morphology, loss or degeneration of tissues, and lethality. These phenotypes were the results of defective actin filament formation and Datx2 was found to be a dosage-sensitive regulator of actin filament formation. Since loss of cytoskeleton-dependent dendritic structure reveals an early event
in SCA2 pathogenesis, the study suggested the possibility that dysregulation of actin cytoskeletal structure resulting from altered ataxin-2 activity is responsible for neurodegeneration in SCA2. Additionally, results also demonstrated that proper Datx2 expression levels are indispensable for regulating the morphology of actin structures in photoreceptor cells.

Ataxin-2 binding protein (A2BP1) that binds to the C-terminus of Ataxin-2 was isolated by yeast-two hybrid screening (Shibata et al., 2000). As the C-terminal Ataxin-2 fragment does not include the poly(Q) tract, A2BP1 does not require the poly(Q) tract for binding. A2BP1 is predominantly expressed in muscle and brain. Both A2BP1 and Ataxin-2 were localized to the trans-golgi network. It was also found that, A2BP1 was also expressed in the cytoplasm of Purkinje cells and dentate neurons as that of Ataxin-2. The specific binding of A2BP1 to Ataxin-2 in brain might contribute to the restricted pathology of SCA2. A2BP1 has an RNP motif that is highly conserved among RNA-binding proteins. Another study carried out by the same group confirmed that the SCA2 gene product, Ataxin-2, was predominantly located in the Golgi apparatus. The deletion of ER-exit and trans-Golgi signals in Ataxin-2 resulted in an altered sub-cellular distribution and the expression of full-length Ataxin-2 with an expanded repeat disrupted the normal morphology of the Golgi complex (Huynh et al., 2003). In yeast, PBP1 and Pab1p that are similar to proteins, Ataxin-2 and A2BP1 respectively, are known to regulate polyadenylation (Mangus et al., 1998; Shibata et al., 2000).

Mouse ortholog of A2BP1 with RNA binding motifs has been characterized and it shares high degree of homology at sequence level to other A2BPs of human, C.elegans and Drosophila (Kiehl et al., 2001). It was suggested that these proteins have a major role in development and differentiation. Earlier RNA interference has been used in C.elegans to characterize the function of worm orthologs of Ataxin-2 and A2BP1, named atx-2 and fox-1 respectively and found that these genes play an important role in early embryonic development. Fox-1 is a known sex determinant gene functions as a numerator element that reduces the effect of the extra X chromosome (Chr) in the hermaphrodite to a haploid dose effect (Hodgkin et al., 1994).
The present study is aimed at functional characterization of Drosophila homologue of Ataxin-2 binding protein, dA2BP1 encoded by CG32062. CG32062 contains both PolyQ stretches and an RRM-type RNA-binding domain.

3.2. RESULTS

3.2.1. CG32062-Transcripts and Polypeptides

CG32062 (Flybase ID: FBgn0052062) is located on Chromosome 3L and cytologically mapped to location 67E4-67E5. Its sequence location is 3L: 10509105...10586521. It is 77.42kb long coding for annotated 5 transcripts and 5 polypeptides (Fig 3.2). Its second intron is 50kb-long, with number of regulatory sequences and other binding sites present in it. The five isoforms are CG32062-RB, CG32062-RD, CG32062-RE, CG32062-RF and CG32062-RG. There are 2 genomic clones and 65 cDNA clones available for this gene. Among them four of the cDNA clones are fully sequenced and 13 of them are end sequenced. Based on the analyses of ESTs for this gene (http://flybase.bio.indiana.edu/), as many as 17 exons have been predicted with at least 5 transcripts (coding 5 protein isoforms), with only 5 exons being common to all isoforms. Among the four fully sequenced cDNAs, LD15974 is for the isoform CG32062-RE while AT08247 represents the isoform CG32062-RF. The signature feature of the sequence is the RNA-binding domain, present in the middle of the sequence coding for 92 amino acids. The RBD domain is of RRM type containing the two conserved RNP motifs, RNP1 and RNP2. Protein sequence of CG32062-RE shows stretches of PolyQ (16 No) in N terminal and dispersed stretches of PolyQ in C terminal (Fig 3.3 and Fig 3.4).
Fig 3.2 Schematic of CG32062 as shown by FlyBase. CG32062 has 5 transcripts and 5 isoforms. The isoform CG32062-RE is of interest to the present study.
Fig. 3.3 Schematic and protein sequence of the isoform CG32062-RE. Stretches of PolyQ both in the N-terminus as well in the C-terminus is seen. Underlined protein sequence in the middle of the protein represents the RBD domain.

Fig. 3.4 Cartoon of CG32062 showing PolyQ and RBD domain. CG32062-RE is of 3457 bp. Open Reading Frame (ORF) is of 2594bp coding for 866 aa representing protein of 94kDa while the RBD domain
3.2.2. CG32062 is the *Drosophila* homolog of A2BP1 group of proteins, sharing RNA binding motifs.

Different transcripts and protein sizes were found in CG32062, similar to human and mouse isoforms. CG32062 encodes for an RNA-binding motif (amino acids 531-615) of 84 amino acids. By database searching, it was found that CG32062 is homologous to human A2BP1 (P83A20), *Mus musculus* A2bp1, *C. elegans* homologs (T07D1.4, R74.1) and *Ornithorhynchus* oA2BP1. Mainly at the RNA-binding domain, it had very high homology to those sequences. When CG32062 protein sequence was compared with sequences of other insect groups like Anopheles, *Apis mellifera*, it was found that it shows very high conservation mostly at the RBD level. The entire RNP region amino acids 531-615 of CG32062 showed an extremely high conservation. Seventy-six out of 84 amino acids were identical in all of these proteins. The RNA binding motif present in CG32062 is of RRM-type containing two-consensus sequence RNP-1 and RNP-2. The consensus sequences, RNP-1 (K-G-F-G-F-V-T-F/M) and RNP-2 (L-H-V-S-N-I) were common to CG32062, vertebral A2BP1 groups and other known insect A2BP1 groups. This strongly suggests that CG32062 is a member of an A2BP1 gene family. As mentioned above, RNA-binding proteins with RNP motifs are involved in various RNA-metabolic functions such as RNA processing and intracellular RNA transport. The evolutionary conservation suggests that protein in this family may be involved in critical molecular events. (Clustal Analysis 1)

3.2.3. Expression pattern of CG32062

To study the expression pattern of CG32062 in detail, antibodies against CG32062 was raised in rabbit. The protein encoding the entire ORF of the gene *CG32062* was cloned from the cDNA encompassing the full-length sequence (LD15974 that codes for the isoform CG32062-RE) into the expression vector pET21a and overexpressed in the bacterial system. Over expressed protein was not in the supernatant but in the inclusion body (Fig.3.5A and B). Nevertheless, the over expressed protein was of correct molecular
Cluster Analysis. CG32062 is homologous to human A2BP1, Mus musculus \textit{A2bp1}, and \textit{Ornithorhynchus} \textit{A2bp1}. Mainly at the RNA-binding domain, it had very high homology to those sequences.
Fig. 3.5 Induction of CG32062 protein in the bacterial system. On induction after 4 hours, (A) Thin band of 94 kDa is seen in the soluble fraction (B) while a broad band of 94 kDa is seen in the inclusion bodies. (C) Western Blot analysis of the bacterial protein using anti-His antibody. Figure shows a specific single band of 94 kDa (lane 2). (D) An attempt to purify the protein using the His-tag resulted only in low yield.

UI = Uninduced, I = Induced.
weight. It was further confirmed by Western blot analysis using antibodies against the His-tag (Fig 3.5C). Attempts to solubilize the over expressed protein and purifying from the tagged column failed (Fig.3.5 D). Therefore, protein extracted from the SDS-PAGE was used for raising antibodies. A week before injecting CG32062, the pre-immune serum was collected from the rabbit. The gel-stripe of the over expressed protein in the inclusion body was crushed and injected into the rabbit for raising antibodies. Three times, at an interval of 15 days, three booster injections were given. Antiserum was collected from the rabbit and tested for activity. Antiserum was further purified against the over expressed bacterial protein separated on SDS-PAGE (antibodies binding to the protein in the gel-stripe itself). It showed a single band of right molecular weight (94kDa) on Western blot analysis against the over expressed protein (Fig 3.6A ). Using the antibody, the distribution of CG32062 protein in different parts of the *Drosophila* tissues was investigated both by Western blot analyses as well as by immunohistochemistry.

Western blot analysis on wildtype embryos, third instar whole larval tissue and dissected wing discs, detected a single band of 94kDa in embryos and from wing discs suggesting that CG32062-RE is the predominant isoform (Fig 3.6 B).

Immunocytochemistry complemented the results obtained by Western blot analysis. Specific pattern of CG32062 expression is first observed in retracted germ band stage, mainly in central nervous system and peripheral nervous system (Fig 3.7 (A'-A” and B'-B”)). It is also expressed in the pericardial cells. In post-embryonic stages, CG32026 is observed in all imaginal discs. In eye discs, anti-CG32062 stained a subset of cells in morphogenetic furrows (Fig 3.7 (C'-C”)). In leg and antennal discs, it detected a concentric ring pattern at the center (Fig 3.7. (D'-D”)). In haltere discs, CG32062 was found to be restricted to the posterior side of the non-D/V cells (Fig 3.7(E'-E”)). In wing imaginal discs, CG32062 marked the non-Dorso/Ventral (non-D/V) cells on either side of the D/V boundary. It is also expressed at very high levels in the notum region of the wing disc and in peripodial cells (Fig.3.7(F'-F”)).

In third instar wing and haltere imaginal discs, the expression pattern was further analyzed using specific markers. Expression pattern of homeodomain protein Cut marks
Fig. 3.6 Western Blot analysis using anti-CG32062 antibodies. (A) Different amounts of protein induced in bacterial system showing a specific single band of 94kDa. (B) Proteins extracted from various Drosophila tissues. A specific band of 94kDa is seen for embryonic tissue (Lane 7), similar to that of the bacterially expressed protein (Lane 10), only when affinity purified antibodies were used.

Emb = Embryonic tissue, ED = Wing discs, WLT = 3rd instar whole larval tissue, + = bacterially expressed protein.
Fig. 3.7. Expression pattern of CG32062 in different tissues of *Drosophila*. (A-B) Embryos, (C) CNS, (D) Eye disc, (E) Leg disc, (F) Haltere disc and (G-H) Wing disc.
the D/V boundary and the myoblasts of the notum region of the wing disc. Double staining using the antibodies against CG32062 and antibodies against Cut showed a clear expression of CG32062 in the non-D/V cells (Fig 3.8A'). While in the notum region both these proteins colocalized suggesting that CG32062 is expressed in the myoblasts of the notum (Fig 3.8A'”). Similarly double staining using antibodies against Ubx, which marks the peripodial cells of the epithelial sheath in wing disc and CG32062 showed a clear colocalisation in peripodial cells (Fig 3.8B). At the sub-cellular level, CG32062 is predominantly a nuclear protein.

3.2.4. Comparison of expression patterns of CG32062 and EN403-lacZ

*EN403* is an enhancer trap that was identified as one of the novel markers and a potential target of Homeotic gene control (Bajpai et al., 2004). This enhancer line is inserted in the II intron of the gene CG32062. *EN403*, enhancer-trap of CG32062, is differentially expressed between wing and haltere imaginal discs. For further studies on the regulation of CG32062 expression, a lacZ enhancer trap was generated by targeted P-conversion approach (Sepp and Auld, 1999). RNA in situ using LD15974, an EST mapped to CG32062 was also carried out. Both EN403-lacZ pattern and RNA in situ pattern showed an expression pattern identical to that of EN403-GAL4 in wing discs. These observations suggest that *EN403* is an enhancer trap of CG32062 (Bajpai et al., 2004).

Now in the present study too, a detailed comparative study with this available marker EN403-lacZ was carried to characterize CG32062. In the embryo EN403 is expressed in the CNS, although no specific pattern could be recognized. CG32062 is also expressed in the embryonic CNS (Fig.3.9(A’-A’”)). In eye discs too, they show similar expression pattern: they both are expressed in the morphogenetic furrow (Fig 3.9(B’-B’”)). Both EN403 and CG32062 are expressed in a subset of non-D/V cells of the pouch, on either side of the D/V boundary (Fig 3.9(C’-C”’)). In the peripodial cells of wing imaginal discs, both EN403 and CG32062 are expressed. Similar to EN403, CG30262 is differentially expressed between wing and haltere discs. In wing discs, both EN403 and
Fig. 3.8. Expression pattern of CG32062 in Drosophila Wing Imaginal Disc. CG32062 is expressed in non-D/V cells and the myoblasts of the notum region of the wing disc. (A'-A") Anti-Cut (green) marks the D/V boundary as well as the myoblasts in the notum region. Double staining with anti-CG32062 and anti-Cut shows the expression of CG32062 in the non-D/V cells (A') and in the notum region of the wing disc (A". (B'-B") CG32062 is expressed in the peripodial cells of the wing disc stained with anti-CG32062 (Red) and Anti-Ubx (green), which marks the peripodial cells. Merged image confirms the same.
Fig. 3.9. Comparison of expression patterns of CG32062 and EN403-lacZ. Both are expressed in the CNS (A'-A") and in the eye disc (B'-B"). Both show a similar pattern in non-D/V cells of the wing disc (C). Both are expressed only in the posterior compartment of the haltere disc (D'-D") and in a concentric pattern in leg discs (E'-E")
CG32062 are expressed in both anterior and posterior compartments, whereas in haltere discs they are expressed only in the posterior compartment (Fig 3.9(D’-D’')). Similar to EN403, which shows an expression in the presumptive distal segments of leg and antennal discs, CG32062 is expressed in pattern of concentric rings in the presumptive distal region (Fig 3.9 (E’-E’’)). All these results reinforce the fact that EN403 is a real enhancer trap of the gene CG32062.

Double staining for EN403-lacZ and CG32062 suggested complete overlap of EN403-lacZ with that of CG32062, while the reverse was not true. For example, in wing imaginal discs, unlike EN403-lacZ, CG32062 also marks the notum region of the wing disc. This suggests that EN403-lacZ only partially recapitulates the expression pattern of CG32062.

3.3. DISCUSSION

The gene CG32062 that was identified by enhancer trap method (Bajpai et al., 2004) codes for five transcripts corresponding to five protein isoforms. The peptides were aligned and analyzed among themselves. In general, CG32062 possess stretches of polyQ both in the N-terminal and C-terminal region and a RNA-binding domain of RRM type separating these polyQ stretches. Isoforms of CG32062 show a major variation at the C-terminal region rather than at the N-terminal region. Other than the isoform CG32062-RG, at RBD domain, they are 100% conserved. In CG32062-RG it is only conserved partly.

On comparing this sequence with sequences of other species, it was found that at RBD domain level it bears 70% homology to Ataxin 2 Binding Protein 1 (A2BP1). All the members of A2BP1 group of proteins are characterized by the presence of RNA binding domain of RRM type with the consensus RNP1 and RNP2 motifs in them. A ClustalW Multiple Sequence Alignment was done against CG32062 isoforms with that of human Ataxin 2 Binding protein (hA2BP1), mouse Ataxin 2 Binding protein (mA2bp1), Anopheles and Apis mellifera protein sequence. It is found that CG32062 is conserved to
a greater extent with hA2BP1 and mA2bp1, but only at the RNA-binding protein level. As expected, RNA binding protein domain shows homology to a greater extent to *Anopheles* than to *Apis mellifera* sequence. In the absence of any other similar proteins in flies and as such there is no conservation at other regions amongst these sequences, it appears that CG32062 is the *Drosophila* homolog of vertebrate A2BP1 proteins.

In CG32062, as there are two different domains, namely the RBD domain in the middle and polyQ domains both in the N-terminal and C-terminal ends, two different functions can be speculated. First, as CG32062 contains RRM domain, it can interact with RNA molecules in its functional domain. Second specific interactions with other proteins can be accounted by polyQ domains. Or it is also possible that RNPI and RNP2 of RBD domains bring about required confirmation change, thereby recruiting other proteins to be regulated by polyQ domain or vice versa wherein polyQ aggregates different proteins whereas RBD domain regulate their activity.

In the present study, to characterize the *Drosophila* homolog of A2BP1, antibodies against CG32062 were raised using the cDNA available for the isoform CG32062-RE. Antibody was used to determine the expression pattern of CG32062 in wing imaginal discs and other tissues of *Drosophila*. Regardless of tissue or cell types, CG32062 is predominantly detected in the nucleus, with very low levels of expression in the cytoplasm. Interestingly, in contrast to most of the other proteins implicated in polyglutamine disease, human and mouse Ataxin-2 and A2BP1 are cytoplasmic rather than nuclear proteins.

Expression pattern analysis is the first step to fully understand and characterize the gene function. From the expression pattern analysis, it appears that this gene might have different roles in different tissues including nervous system and muscle development. Genetic analysis of the function of CG32062 has been described in the subsequent chapters.