DNA binding specificity and transcriptional regulation of Six4, a myotonic dystrophy associated transcription factor

Theodore Kiosses

A thesis submitted for the degree of PhD

Institute of Cell Biology

University of Edinburgh

2008

Table of Contents

Chapter 1 – Introduction

Foreword

Attaining an understanding of the mechanisms underpinning development has been amongst the cardinal scientific challenges of our age. The transition from a single cell organism to the level of complexity evidenced in higher eukaryotes has been facilitated by the advent of intricate developmental networks involving a plethora of factors that synergise to allow for precise spatio-temporal expression of the proteins present in higher organisms. Development is often portrayed as a domino-like cascade of events stemming from relatively uncomplicated origins that go on to branch out and form associations and interactions amongst multitudinous actors that will inexorably lead towards a higher state of order. Transcription factors occupy a central position within this tapestry of interactions. They regulate expression of the various required proteins and they provide the cues for the developmental events that will eventually shape an organism. These factors frequently remain unknown until some occurrence causes developmental processes to fail and inadvertently focus attention on the factors that facilitate development. Myotonic dystrophy is a useful paradigm of such a developmental dysfunction that has led to the discovery of a transcription factor integral to both muscle development and gonadogenesis in both *Drosophila* and higher eukaryotes.

1.1 Myotonic Dystrophy Overview

Myotonic dystrophy (DYSTROPHIA MYOTONICA 1, Steinert disease or DM1, OMIM 160900) is an autosomal dominant disorder characterized by ‘‘myotonia, muscular dystrophy, insulin resistance, cataracts, hypogonadism, frontal balding, ECG changes (heart conduction defects) and mental disturbances’’ (Harper, 2001). Many of the clinical features of this disorder are attributed to an amplified trinucleotide repeat (CTG) on chromosome 19 in the 3-prime untranslated region of the *DMPK* gene, a serine/threonine protein kinase (Mahadevan et al., 1992; Harley et al., 1992; Aslanidis et al., 1992; Brook et al., 1992; Wheeler and Thornton, 2007), and also in the promoter of the *SIX5* homeodomein protein gene (Klesert et al., 1997). DM1 is known to cause pathogenesis through several potentially interacting pathways with *SIX5* haploinsufficiency being the one of most interest to this study. Attaining a complete understanding of the pathology of DM1 is important if one is to decode the developmental networks that it affects and segregate the developmental defects attributed to the dysfunction of the SIX5 protein caused by the RNA mediated effects

of repeat expansion. The various aspects of the DM1 aetiology as well as its pathophysiology are presented in greater detail in the following sections.

1.2 DM1 Phenotype

Instances of DM1 are divided into adult onset and congenital cases. The most prominent clinical features of adult onset myotonic dystrophy are myotonia, muscle degeneration and weakness, ocular cataracts and hypogonadism. The organ system most severely affected by DM1 is the skeletal muscle. Muscles affected by DM1 include the distal muscles of the extremities as well as the proximal musculature, and the muscles of the head and neck. Myotonia, delayed muscular relaxation following contraction, is frequently apparent in the tongue, forearm, and hand (Harper, 2001). Additional features of adult onset DM1 also include facial disfigurement and kidney failure (Barbosa et al., 1974).

Another developmental symptom is the impaired responsiveness to follicle stimulating hormone with hypogonadism and impairment of adrenal androgens, and occasional thyroid dysfunction (Sagel et al., 1975; Sarkar et al., 2004).

DM1 symptoms also include widespread nervous system dysfunction (Jamal et al., 1986) and IQ decline, probably caused by white matter hyper-intense lesions in the brain (Di Costanzo et al., 2008; Turnpenny et al., 1994), as well as significantly increased cortical atrophy (Censori et al., 1994) and cardiac pathology features (Togkozoglu et al., 1995; Bu'Lock et al., 1999) including cardiac autonomic nervous system dysfunction (Rakocevic-Stojanovic et al., 2007). Diabetes mellitus is observed in 5% of cases, frequently with hypersecretion of insulin, suggesting insulin resistance (Ristow, 2004; Barbosa et al., 1974).

Congenital myotonic dystrophy (Myotonia Congenita) constitutes only a small fraction of DM1 cases and its symptoms include neonatal hypotonia, motor and mental retardation, and facial diplegia as well as frequent and often fatal respiratory difficulties (Harper, 2001). Additionally congenital DM1 patients display most of the symptoms described above, often with increased severity.

1.3 Myotonic Dystrophy Genetics

The cause of myotonic dystrophy has been localized to chromosome 19 (Mahadevan et al., 1992). DM1 is a genetically inherited autosomal dominant disorder of variable penetrance with many carriers being asymptomatic. Homozygotes are not

more severely affected than heterozygotes and thus DM1 is a ‘‘true dominant condition’’ (Cobo et al., 1993). In cases of Myotonia Congenita the disorder is almost exclusively transmitted by the mother with paternally transmitted cases being less severe (Harper and Dyken, 1972). The maternal effect on age of onset and severity (Andrews and Wilson, 1992; Tanaka et al., 2000) is attributed to the chemical factor deoxycholic acid. It had been proposed that abnormalities of bile acid metabolism play a pathogenetic role in DM1, in which deoxycholic acid acts as a maternal factor in association with the onset of congenital DM1 (Tanaka et al., 1981; Tanaka, 1985). Additionally the expression of the *DMPK* gene was not found to be subject to imprinting or mitochondrial genetic modification (Jansen et al., 1993). The true causes of DM1 however can be found on the molecular level.

1.4 Molecular Genetics of DM1

As mentioned in section 1.1, DM1 is caused by an expansion of the CTG repeat region in the 3-prime untranslated region of the *DMPK* gene (Harley et al., 1992) (Fig. 1.1), a gene which encodes multiple protein isoforms of a serine/threonine protein kinase (Mahadevan et al., 1992**)** and in the promoter region of the immediately adjacent *SIX5* homeodomain gene. There is a positive correlation between the size of the repeat and the age of onset and severity of the symptoms of the pathology. The number of repeats is highly unstable and the repeat regions themselves are subject to expansion. DNA mismatch repair proteins, Msh2 and Msh3, are required for the formation of intergenerational and somatic expansions (Foiry et al., 2006).

The expanded CTG repeats are known to interfere with the activity of Muscleblind-like proteins (MBLP) through a mechanism of RNA-mediated toxicity. The function of this mechanism has been demonstrated through the introduction of expanded CTG repeats in both *Drosophila* (Garcia-Lopez et al., 2008) and murine models (Mankodi et al., 2000; Orengo and Aguade, 2007; Gomez-Pereira et al., 2007; Tapscott, 2000). These methods have succeeded in reproducing most of the DM1 phenotype in these systems.

This has shifted the focus of explanation of DM1 aetiology from the originally proposed reduction in active *DMPK* (and *SIX5*) transcripts to RNA-mediated toxicity. The later mechanism however fails to account for the entirety of the phenotype and knock-out experiments (see below) involving these proteins have shown their involvement in DM1 pathology (Wang, 2007). Chromatin disruption caused by the

CTG expansion has been known to cause alterations in protein expression linked to the DM1 phenotype. Chromatin structure is altered through CpG methylation, histone modifications, chromatin remodelling factors, and non-coding RNA. The roles of both DMPK and SIX5 in development have since been investigated in model organisms. Specifically, knockout experiments suggest the involvement of SIX5 in a number of developmental processes. The function of SIX5 is outlined in the following sections.

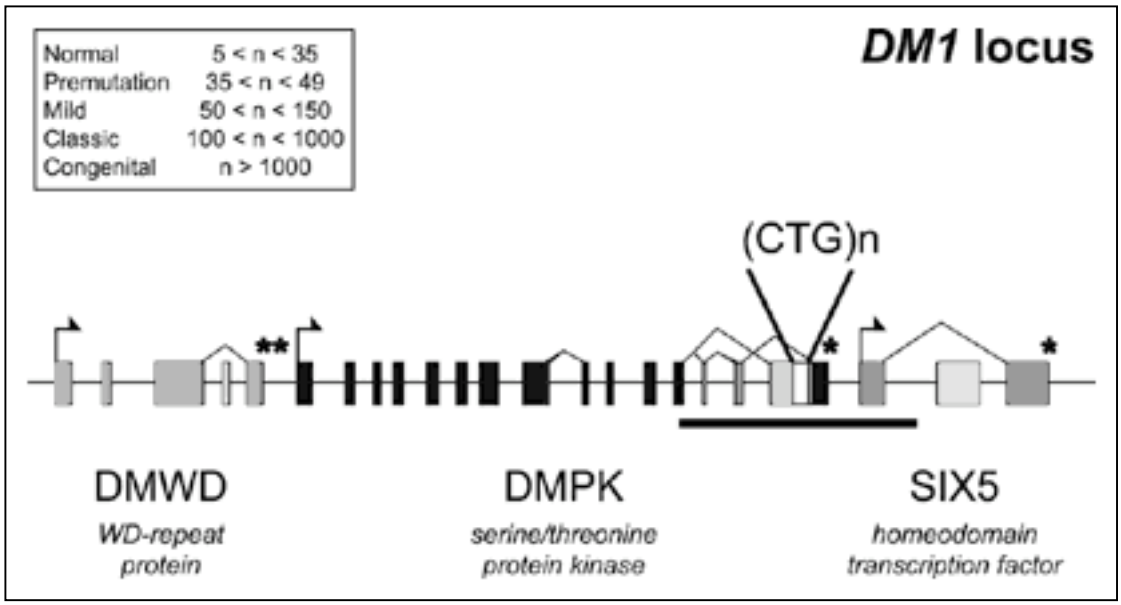


Fig. 1.1 Schematic representation of the DM1 locus: The DM1 locus contains three genes, *DMWD*, *DMPK* and *SIX5*. *DMWD* encodes a WD-repeat protein, *DMPK* a serine/threonine protein kinase, and *SIX5* a homeodomain transcription factor (section 1.5). Transcription initiation sites are indicated with arrows; exons with boxes, and introns and intergenic sequences with a straight line; alternative splice modes with connecting lines between exons, and polyadenylation sites with asterisks (two for *DMWD*). The unstable (CTG)n repeat is located in exon 15 of the human *DMPK* gene, in contrast, the mouse locus contains a (CTG)2(CAG)2(CTG)-sequence. The length of the repeat is variable, and is strongly correlated with age of onset and severity of DM1 disease symptoms. (Figure presented as published in Wansink and Wieringa, 2003)

1.5 The involvement of DMAHP (SIX 5) in DM1

The first suggestion that DM1 pathophysiology may be due to a malfunction of more genes than just *DMPK* came when it was shown that strong nucleosome positioning signals were created by CTG repeats when nucleosomes were reconstituted on DNA *in vitro* (Wang et al., 1994). However, *DMPK* knockout mice or in mice that over-express a human *DMPK* transgene showed only minor histopathologic abnormalities. This was found to be in contrast to the widespread muscle wasting that characterises DM1 making it unlikely that changes in *DMPK* are the sole cause of the disease (Thornton et al., 1997). In light of the recent findings on RNA-mediated toxicity it is apparent that the DM1 phenotype isn't caused by altered *DMPK* transcript levels. Prior to these findings, however, the involvement of additional genes seemed likely and the genomic vicinity of *DMPK* was searched for phylogenetically footprinted areas. The area 3' of the *DMPK* locus was found to contain a region of conservation between the human and mouse genomic sequences hinting towards the presence of another gene near *DMPK* (Boucher et al., 1995). This led to the identification of the DM1 locus-associated homeodomain protein (DMAHP) or SIX5. SIX5 is a transcription factor that belongs to the homeodomain family and has been shown through RT-PCR analysis to be expressed in a number of human tissues, including skeletal muscle, the brain, and the heart (Korade-Mirnics et al., 1999). Other genes known to be differentially expressed in DM1 patients include *DMWD*, which is situated immediately upstream (500bp) of *DMPK* (Fig 1.1) that has been shown to be strongly expressed in the brain and testis (Shaw et al., 1993) and genes *GIPR* (human gastric inhibitory polypeptide receptor gene), *symplekin* and *20-D7* (Alwazzan et al., 1998)

Sequence comparisons pointed towards a homology between *DMAHP* and the *Drosophila* eye development gene *sine oculis* (*so*) which led to *DMAHP* being renamed *SIX5* (after *sine oculis*). Two splicing isoforms of *SIX5* mRNA have been isolated although no protein product of the shorter splicing isoform of *SIX5* (which has an altered carboxy-terminus) has been detected (Pham et al. 2005). Two binding sites of the zinc-finger protein CTCF flank the *SIX5* locus. They insulate the locus by blocking enhancer promoter communication necessary for independent transcription regulation of *SIX5* and *DMPK* (Filippova et al., 2001).

Since the discovery of *SIX5*, it has been shown that steady-state *SIX5* transcript levels in cells of myotonic dystrophy patients show a 2- to 4-fold reduction relative to

wildtype controls (Klesert et al., 1997). It was similarly shown through allele specific quantification of SIX5 expression using an RFLP within *SIX5*, that the DM1 mutation reduces SIX5 expression in myoblasts, muscle, and myocardium. The severity of this effect is partially proportional to the extent of the CTG repeat expansion. These findings established a link between SIX5 and the pathophysiology of DM1 (Thornton et al., 1997).

The homology of *SIX5* to *sine oculis* hinted towards SIX5 being responsible for the ophthalmic features of DM1. This suspicion has been confirmed by findings of Pham et al. (2005) showing expression of SIX5 in adult eyes. Additionally, Winchester et al. (1999) reported ‘‘a restricted but partially overlapping expression pattern for DMPK transcripts and DMPK protein in normal foetal and adult eyes’’, thus strengthening the case that it was SIX5 and not DMPK that was responsible for the development of adult-onset cataracts, the most frequently occurring eye phenotype in DM1. Additionally, SIX5 has been shown to be mutated in patients with Branchio-oto-renal syndrome (BOR) an autosomal dominant disorder characterised by branchial arch defects, hearing loss, and renal anomalies (Hoskins et al., 2007). Mutations in EYA1, a known co-factor of SIX5, are known to cause similar symptoms. These findings are consistent with the formation of an EYA1-SIX5 complex to activate gene transcription. Attempts to elucidate the function of SIX5 and DMPK, and their contribution to DM1 pathology, have been made through disruption of their murine orthologues, Six5 and Dm15.

1.6 SIX5 involvement in DM1 and the disruption of the murine Six5

A targeted deletion of *Dm15* (also known as *dmpk*), the mouse orthologue of *DMPK*, in a murine model (Reddy et al., 1996; Berul et al., 1999) produced mice with a mild myopathy and heart conduction abnormalities but lacking the other symptoms of DM1 such as cataracts and myotonia.

Similarly, disruption of the mouse *Six5* (orthologue of the human *SIX5*) results in homozygous mutant mice that exhibit no skeletal muscle defects, but develop lenticular opacities at an increased rate. This cataract development is temporally progressive and occurs in a dosage dependent manner with homozygotes being more severely affected than heterozygotes (Klesert et al., 2000; Sarkar et al., 2000). Both the severity and latency of cataract formation show variable penetrance. The progressive destruction of lens tissue caused by loss of Six5 is not entirely consistent

with homologous mutations in *Drosophila* however. Loss of Sine oculis, the fly orthologue of Six3 (a protein belonging to the same family as Six5, see section 1.9) causes massive cell death anterior to the morphogenetic furrow resulting in the disruption of the entire visual system in flies (Cheyette et al., 1994). However, loss of Six4, a *Drosophila* homologue of Six5, causes a number of defects that do not have any impact on eye development (Kirby et al., 2001).

Both homozygous and heterozygous mice show increased steady-state levels of the Na+/K+-ATPase -1 subunit (ARE) and decreased *Dm15* mRNA levels. It has subsequently been shown that ARE is a direct regulation target of Six5 (Kawakami et al., 1996). It is unclear if the altered *Dm15* mRNA level is a consequence of the loss of Six5 function or whether it is due to the *cis* effects associated with the targeted deletion of *Six5*.

In addition, data by Sarkar et al. (2004), demonstrates ‘‘a strict requirement of Six5 for both spermatogenic cell survival and spermiogenesis in mice with Leydig cell hyperproliferation and increased intra-testicular testosterone levels being observed in the *Six5-/-* mice and steady-state c-Kit levels being reduced in the *Six5-/-* testis’’. The decreased c-Kit levels provide a likely explanation for spermatogenic cell apoptosis and Leydig cell hyperproliferation in the *Six5-/-* mice. These findings establish a link between the reduced levels of SIX5 and the male reproductive defects in DM1 (Sarkar et al., 2004).

Finally, *Six5* heterozygous mutants were shown to have heart conduction abnormalities, particularly infraHisian conduction delay, one of the initial phenotypes of adult-onset cardiac conduction abnormalities in DM1 patients (Wakimoto et al., 2002). However, loss of Six5 function in mice does not affect viability.

The above observations suggest that the cataract phenotype and reproductive defects of DM1 can be attributed to SIX5 deficiency and that the rest of the pathology is caused by changes to levels of transcription caused by the epigenetic effects of CTG expansion. The boundaries to which such epigenetic effects spread as a function of repeat tract length may provide insights into the multisystemic nature of the DM1 phenotype. Finally, an essential step in elucidating the role of SIX5 is deciphering its role as a transcription factor and the implications of its absence.

1.7 Transcription factors

The various processes that underpin development are largely mediated through the enactment of tightly regulated programs. The functional units of these programs are contained within *cis* regulatory elements or modules (CREs or CRMs) that are usually located in the vicinity of the genes they regulate. It is through the function of these CRMs that developmental control is maintained. *Cis*-regulatory target sites recruit transcription factors required to control the expression of the genes associated with them in a sequence specific manner. These target sites control the docking of transcription factors (and through them the basal transcription apparatus). These factors, and other proteins that in turn bind to them (known as co-factors), determine the rate of transcription and mediate the accurate activation or repression of the gene in a precise spatio-temporal manner thus determining the geography of a developing organism through the specification of different cell types and lineages. ‘‘The identities of the genes encoding those transcription factors that, in terms of causality, lie directly upstream of any given *cis*-regulatory system are therefore determined by its target sites’’ (Arnone and Davidson, 1997). SIX5 and its homologues function in this way and share the characteristics of most transcription factors as well as those of members of the homeodomain transcription factor family they are members of.

1.8 Homeodomain transcription factors

Homeodomain genes constitute a distinctive category of transcription factors that are named after their characteristic DNA-binding motif. The homeodomain constitutes one of the most studied eukaryotic DNA-binding motifs. It was discovered when homeotic mutations, i.e. mutations leading to segmental transformations, were observed in *Drosophila* (Gehring, 1966; Lewis, 1978) and later localized in genes encoding a stable domain of about 60 residues (Chi, 2005; McGinnis et al., 1984; Scott and Weiner, 1984). DNA-binding is usually controlled through the 50^{th} homeodomain position which is often occupied by Glutamine. Homeodomain proteins are common to various species and regulate numerous developmental processes, often in an interspecifically analogous manner (Duboule and Morata, 1994). Processes controlled by homeodomain factors include regional specification, patterning, migration and differentiation (Gehring et al., 1994_1).

Homeodomain structure is well studied and forms the basis of homeotic gene classification. A typical homeodomain is composed of three helices, which are folded around a hydrophobic core in which the second and third helix adopt a helix-turn-helix motif for DNA recognition, and a flexible N-terminal arm with additional important functional roles (Gehring et al., 1994_2; Billeter et al., 1996; Wolberger, 1996). The third helix and the N-terminal arm recognize the major groove and the adjacent minor groove of target DNAs, respectively. The N-terminal arm also contains a stretch of basic residues known as the nuclear localization signal (NLS). Unlike conventional helix-turn-helix motifs, which use the residues on the turn and the first loop of the third helix to contact DNA, homeodomains make these contacts with residues that are located toward the C-terminal end of the third helix. This structure shows remarkable conservation between highly different homeodomain factors. Homeodomains are either found alone as a DNA-binding motif or in conjunction with other domains, such as paired-homeodomains (Wilson et al., 1995; Chi, 2005), LIM-homeodomains (Hobert and Westphal, 2000), POU-homeodomains (Ryan and Rosenfeld, 1997), or cut-homeodomains (Harada et al., 1994). This notion of paired homeodomains is important for reasons that will become apparent later. It is conceivable that the homeodomain of SIX5 and its homologues shares this property of cooperative DNA binding.

Homeodomain family members show features of structural and functional conservation. Their three-dimensional structures are more conserved when compared to their primary sequences. This establishes the necessity for a conserved architecture in order to facilitate interactions like DNA recognition and protein-protein interactions. ‘‘Some amino acids, such as Trp48, Phe49, Asn51, and Arg53, which are invariant among almost all homeodomains, are essential in maintaining structural integrity and/or making contacts with DNA, whereas other residues vary in order to provide DNA-binding specificity and other protein functions’’ (Chi, 2005).

All homeodomains are capable of binding to the ATTA/TAAT sequence, with other consensus motifs being specific to the different homeobox families. Recognition of these sequences is often mediated through processes that include post-translational modification, protein-protein binding and DNA-binding of cofactors (Garcia-Fernandez, 2005).

1.9 The SIX family of homeodomain genes

The human *SIX5* gene as well as its murine orthologue *Six5* and the closely related *Drosophila* gene *Six4* are all members of the SIX family of homeodomain genes (Fig.1.2). Transcription factors of the SIX family of homeobox genes are vertebrate homologues of the *Drosophila sine oculis* gene, which is required for the development of the *Drosophila* visual system (Wawersik and Maas, 2000). The SIX proteins share two highly conserved regions, a SIX domain and a SIX-type homeodomain, which is located adjacently (Kawakami et al., 2000).

The SIX family homeodomain is very distinctive and shows considerable divergence from typical homeodomains (less than 30% identity) (Cheyette et al., 1994; Oliver et al., 1995; Seo et al., 1999). Members of the SIX family owe their name to another domain called the SIX domain (SD, 116aa according to Seo et al., 1999) which is located near the amino-terminus of the homeodomain and which is known to facilitate protein-protein interactions. Interestingly, in a recent publication, Hu et al. (2008) report the length of the SIX family homeodomain as being 59aa long (a departure from the canonical homeodomain model) and the SIX domain as being 115-121aa long. Findings presented herein suggest that the latter authors are correct in their limitation of the SIX domain in terms of functionality.

Eyes absent (Eya) proteins are known cofactors of some SIX proteins including mouse Six1, Six4 and Six5 as well *Drosophila* proteins Sine oculis and Six4. Members of the SIX and Eya families have been shown to interact both *in vivo* and *in vitro* through the formation of a functional heterodimer (Ohto et al., 1999; Grifone et al., 2005; Hu et al., 2008). *Drosophila* domain swap experiments performed by Hu et al. (2008) have shown the SIX domains to be responsible for conferring DNA-binding specificity through co-factor binding selection and not through direct DNA-binding. The same authors demonstrated that the presence of Eya, the *Drosophila* member of the Eyes absent group of proteins, causes a 12-fold increase in the DNA-binding affinity of Sine oculis, the archetypal SIX protein that is the *Drosophila* homologue of the human SIX1/2 pair.

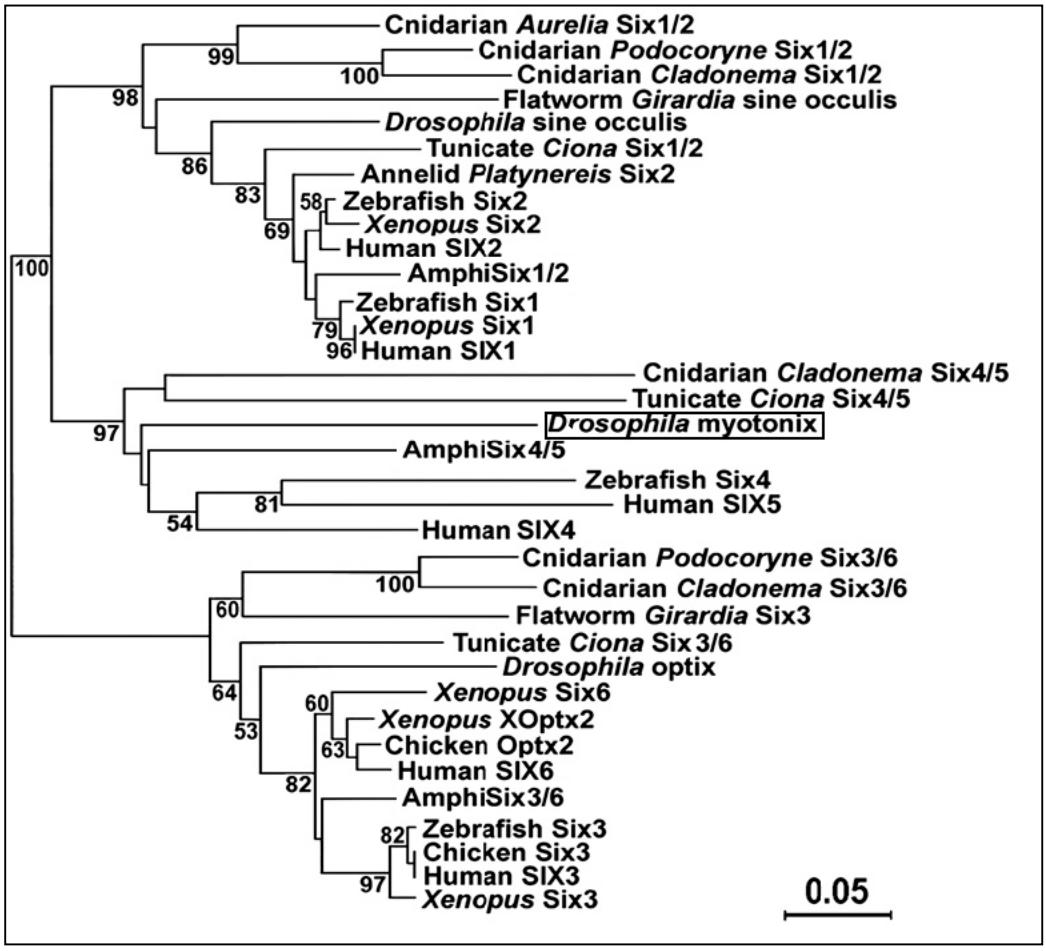


Fig. 1.2 Phylogenetic analysis of the SIX family based on complete animal protein sequence comparisons. The *Drosophila* SIX4/5 class homologue Six4 is presented under its original designation of Myotonix. The tree was constructed by neighbour joining. Node support values are bootstraps (values below 50% have been omitted). The evolutionary distance is calculated as the fraction of aminoacid changes and the scale is included in the bottom right corner. (From Kozmik et al., 2007)

For the *Drosophila* proteins, the amino acids comprising the SIX domain (as defined by Seo et al., 1999), those comprising the homeodomain, and the total number of amino acids are as follows: Six1/2 (7–121; 122–181; 290), Six4/5 (42–156; 157–216; 332), Six3/6 (35–153; 154–213; 250). The topology in Fig.1.2 strongly suggests that all three SIX subfamilies were present in a common urbilaterian ancestor and later separated, respectively, into Six1/Six2; Six3/Six6; and Six4/Six5 within the vertebrate lineage (Jean et al., 1999; Seo et al., 1999; Seimiya and Gehring, 2000).

SIX proteins form part of the *Drosophila* Retinal Determination Gene Network (RDGN) that is a key component of *Drosophila* eye specification. The genes constituting the network (*twin of eyeless, eyeless, sine oculis, eyes absent,* and *dachshund*) are ‘‘interrelated by reciprocal feedback loops and encode nuclear proteins that can form multi-molecular complexes to control target gene transcription’’ (Kozmik et al., 2007). Vertebrate RDGNs, which are involved in such human pathologies as branchio–oto–renal syndrome, which has been directly linked with SIX5 mutation in humans (Sanggaard et al., 2007; Hoskins et al., 2007), comprise over a score of genes belonging to the following four families: ‘‘*Pax* (corresponding to *twin of eyeless* and *eyeless*), *Six*, *eyes absent* (*eya*), and *Dachshund* (*Dach*, unaltered terminology)’’(Kozmik et al., 2007). In a higher eukaryote context the *Drosophila*-specific term “RDGN” is interchangeable with the more widely applicable term “PSEDN” (Pax–Six–Eya–Dach Network) (Kawakami et al., 2000). During vertebrate development, PSEDN genes play key developmental roles not only in the eyes, but also in such structures as muscles, endocrine glands, placodes, and pharyngeal pouches (Hanson, 2001; Silver and Rebay, 2005; Rebay et al., 2005) as well as playing a fundamental role in gonadogenesis (Kirby et al., 2001; Clark et al., 2007). Several members of the So/Six family function as transcriptional activators when interacting with members of the Eyes absent (Eya) gene family (Grifone et al., 2005; Hu et al., 2008). Grifone et al. (2005) report reduced levels of the regulatory factors Myogenin and Myod1, and Mrf4 in *Six1/Six4* double knockout mice. This function however appears to be context specific since Six3 has been shown to act as a repressor of the head developmental protein Wnt1. Additionally Six3 and Six6 have been shown to be involved in the development of the vertebrate visual system (Friedrich et al., 2006). Furthermore, the SIX proteins have been shown to bind the transcriptional co-repressor protein Groucho (Zhu et al., 2002) and are known to be regulated by members of the Groucho family (Lopez-Rios et al., 2003). SIX proteins in *Drosophila* are described in greater detail in the following section (1.10).

Each member of the SIX family is expressed in a spatiotemporally regulated manner during embryogenesis. In mice, *Six3* and *Six6* show expression in the developing forebrain and eyes (Ozaki et al., 2001; Jean et al., 1999; Lopez-Rios et al., 2003; Oliver et al., 1995; Toy and Sundin, 1999), whereas *Six1, Six2*, and *Six5* are expressed in a wider range of tissues (Klesert et al., 2000; Oliver et al., 1995**).** *Six5* shows a broad expression in branchial arches, limb buds, telencephalon, eye, sclerotomes, and cartilages (Klesert et al., 2000). Such distribution suggests that these genes play specific roles in embryogenesis. *Six3* and *Six6* genes have also been shown to be implicated in forebrain and eye organogenesis through overexpression and misexpression experiments (Ozaki et al., 2001; Kobayashi et al., 1998; Loosli et al,. 1999; Oliver et al., 1995; Zuber et al., 1999). Consistently, *SIX3* mutations cause holoprosencephaly, a severe malformation of the brain in humans (Wallis et al., 1999).

Additionally, the expression of Myf5 (a myogenic cell fate determinant responsible for the onset of embryonic skeletal muscle) is reduced in the limb buds of *Six1(-/-)* and *Six1(-/-) ; Six4(-/+)* mouse mutants despite the presence of myogenic progenitor cells (Giordani et al., 2007). Moreover, SIX homeoproteins are implicated in the regulation of the myogenic regulatory genes *MyoD*, *Mrf4*, and *Myogenin* (Sato et al., 2002).

Due to the extensive redundancy between the vertebrate SIX subgroup members however inferences as to their role are harder to make and often necessitate the generation of double mutants. This obstacle can be circumvented through the study of the SIX family members in *Drosophila* where the functions of each pair are performed by a single orthologue.

1.10 SIX genes in *Drosophila*

The SIX family consists of three distinct sub-families which all have one *Drosophila* member. The members of each of these sub-families have the same tetrapeptide near the N-terminus of the homeodomain (HD). In addition to these distinguishing features, the sequence conservation among members of the same family is higher, with generally the same amino acid substitutions relative to the other families. Comparisons of the HDs, varying in sequence identity levels from 45–100% (Fig.1.3 shows a multiple sequence alignment of the murine members of the SIX family), show that *Drosophila Six* genes belong to one of three major sub-families related to *Six2*, *Six3* and *Six4*, respectively (Fig. 1.2). Whereas Sine oculis belongs to

a family of SIX2-like proteins, the *Drosophila* genes *Six3* (*Optix*) and *Six4* (*Myotonix*) encode homologues of vertebrate proteins in the SIX3 and SIX4 families, respectively. The genome of the common ancestor of insects and vertebrates is thus thought to have contained three different *SIX* genes from which the respective families have originated.

HD and SD sequence identity comparisons between *Drosophila* SIX proteins and their vertebrate (murine) homologues show differences in the degree of conservation. These identity comparison ratios are summarised in Table 1.1 HDs are generally the most conserved domains of these homologous proteins, whereas SDs show greater divergence. There is considerable variability between the N- and C-terminal regions of the three *Drosophila* proteins both in terms of length and sequence. This trait is shared by the vertebrate SIX proteins.

Of the three *Drosophila SIX* genes, *Six4* is the least studied. Its homology to *SIX5,* along with other findings, suggests a broad developmental role for *six4*. As such, *Six4* is the primary focus of this study.

	Six1/Six2	Six3	Six4
Sine oculis	HD 95%, SD 84%		
Optix		HD 97%, SD 77%	
Six4			HD 82%, SD 57%

Table 1.1 Comparisons of protein sequence identity between the *Drosophila* members of the SIX family and their murine counterparts. Percentages represent identity of the HD and SD respectively.

Fig. 1.3 Multiple sequence alignment (MSA) of the SIX- and homeodomains of the 6 murine members of the SIX family of transcription factors. Positions marker with a * denote regions that are conserved in the MSA. Sequence conservation is very high between these domains. Multiple sequene alignment was performed using ClustalW2 (default settings). Sequences obtained through the Ensembl genome browser. Topology of SIX- and homeodomains marked as described by Seo et al. (1999).

1.11 Six4, the *Drosophila* homologue of SIX5

The genomic DNA sequence of *Drosophila Six4* (FBgn0027364) can be found on the left arm of chromosome 3 and its cytological map location is 77E6. *Six4* was first amplified through PCR of *Drosophila* larval cDNA using a degenerate primer set derived from the C-termini of the SIX domains (SD) and homeodomains (HD) of several SIX class proteins with the intention of amplifying related sequences in *Drosophila* and took its name from its similarity to mammalian Six4 (Seo et al., 1999). Comparison of the protein sequence of the combined SIX- and homeodomains shows that Six4 is most similar to SIX4 and SIX5 members of the Six4 family (67% and 65% identity, respectively). A hint towards their common ancestry is that all three proteins have valine in homeodomain position 5, which is a potential contributor to DNA binding specificity, whereas all other SIX proteins have serine or threonine. The *Six4* coding region is 1845bp long, including a 456bp leader sequence, an ORF of 1176bp (Fig.1.4), and a 179bp 3'-UTR with a polyadenylation signal followed by a poly(A) tail. The first potential start sites at nucleotide (nt) 103 (taaaATG) and nt 457 (cattATG) for *Six3* and *Six4*, respectively, are both favourable relative to the Cavener consensus sequence for *Drosophila* translation initiation sites (Cavener and Ray, 1991).

Six4 expression in *Drosophila* was first described by means of in situ hybridization by Seo et al. (1999). During embryogenesis *Six4* expression is observed in the developing head region, mesoderm, and the CNS. ''*Six4* is initially expressed in a dorsal patch that straddles the midline between 85%-90% egg length (EL). This patch is wider dorsally (3-4 cells) than towards its lateral edges.''(Seo et al., 1999). Seo et al. (1999) liken the cephalic expression of Six4 to that of So, which is expressed in a dorsal domain of the head region during the blastoderm stage although the report that Six4 is expressed anteriorly with respect to So. Like Optix, Six4 expression is divided into two domains and persists in the dorsal part of the procephalic lobe during gastrulation and germ band elongation.

The role of Six4 in the development of the gonad and the mesodermally derived musculature is controlled through its mesodermal expression. Seo et al. (1999) detected transcripts in mesodermal cells along the entire germ-band by stages 9-10 in a manner consistent with the transient embryonic segmental characteristics at this stage. Clark et al. (2006) speculate that this expression is controlled by Tinman (Tin) a mesoderm specific homeobox transcription factor that is known to regulate

Drosophila Mef2. Indeed the mesodermal expression of Six4 prior to stage 10 was found to coincide with that of Mef2 and is complementary to that of Tin (Clark et al., 2006). This expression becomes limited to the procephalic region by stage 11. By stage 15, additional sites of expression are observed in the ventral cord and gonads (Seo et al., 1999).

After stage 11 mesodermal expression becomes segmental before becoming limited to the somatic gonadal precursors (SGPs) at stage 13. The SGPs (also known as follicle cell precursors) are located in parasegments (PSs) 10–12, and will eventually form the somatic sheath that surrounds the gonad (Van Doren, 2006; Brookman et al., 1992). *Six4* expression in the SGPs persists after they have coalesced with the migrating germ cell precursors (pole cells) to form the gonad. SGPs are essential for gonadal coalescence, and germ cells remain scattered if SGPs are dysfunctional. This failure of the gonad to coalesce is also observed through loss-of-function of Six4, hinting towards a role of Six4 in SGP mediated gonadal coalescence (Kirby et al., 2001).

The expression pattern of Six4 is consistent with its role in mesodermal development. Six4 is required for the development of various mesodermally derived cell types including the SGPs the fat body and the somatic muscles. The presence of Six4 along with its co-factor Eyes absent (Eya) is sufficient for the specification of these cell types (Clark et al., 2006)

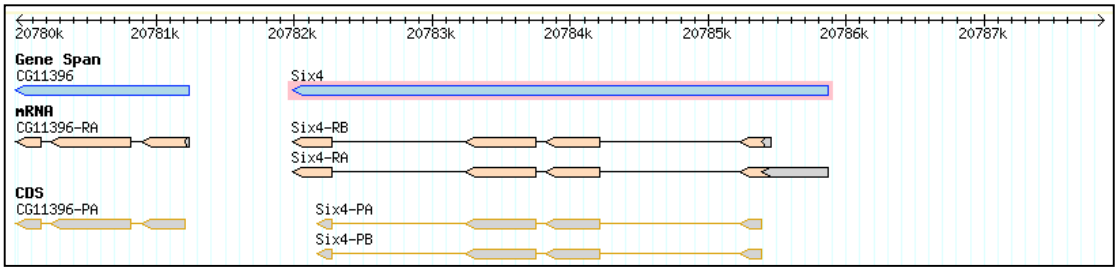


Fig. 1.4. Map of the *Six4* locus (Flybase symbol *Six4* ,CG3871, FBgn0027364). Cytological map location is 77E6. Sequence topology is 3L:20781986..20785868. Expression in the developing *Drosophila* head in regulated through cis-regulatory elements situated within the 4.5kb 5' region directly upstream of exon 1. Mesodermal expression is regulated through elements located within the 3rd intron of the *Six4* gene (see section 4.2). Gene map obtained through FLYBASE (http://flybase.bio.indiana.edu).

1.12 Six4 loss-of-function phenotype

Kirby et al. (2001) injected Embryos with *Six4* dsRNA in order to knock-out Six4 expression. They report that embryos exhibit 100% gonadal coalescence failure. These findings highlighted the requirement for *Six4* in SGPs for gonadal coalescence (Kirby et al., 2001). The same authors isolated two mutations that mapped to the *Six4* chromosomal location and were found through complementation testing to affect the same locus (Fig 1.5). These mutations were designated $Six4^{289}$ and $Six4^{131}$. $Six4^{289}$ is a nonsense mutation (C175 3 > T) that introduces a stop codon in place of Gln87 and results in a truncated protein that lacks the homeo- and SIX domains. Its phenotype is consistent with a complete loss of Six4 function and homozygotes fail to hatch. $Six4^{131}$ is a point mutation (C2404 > T) that results in an amino acid substitution of Cys for Arg281, which corresponds to position 102 within the Six domain. This Arg is conserved in all Six proteins, hinting towards its importance for Six domain structure or function (Kirby et al., 2001). $Six4^{131}$ is less severe and appears to be a milder hypomorph. $Six4^{131}$ mutant embryos hatch normally, although many die during larval and pupal stages with only a small proportion surviving to adulthood.

The $Six4^{289}$ phenotype consistently affects gonadal coalescence, fat body specification and the fusion of mesodermally derived musculature. $Six4^{289}$ homozygotes, undergo initial germ cell internalization and migration but fail at the stage of coalescence in a way reminiscent of SGP failure. Indeed expression of the *412* retrotransposon, an SGP marker, was found to be abolished in $Six4^{289}$ homozygotes by stage 10 with the exception of a few scattered cells that appeared to be SGPs (Kirby et al., 2001). Clark et al. (2006) attribute this phenotype to failure of SGP specification by assaying for the initial presence of Eya, a SGP marker protein. Additionally an abnormally high number of apoptotic cells were detected in the region of the mesoderm normally occupied by the SGPs. These observations are consistent with the requirement of Six4 for gene expression within the mesoderm and SGPs. Additionally Clark et al. (2006) detected very few cells expressing the fat body precursor protein Serpent (Srp) in $Six4^{289}$ mutant embryos by stage 12, thereby implicating Six4 in the specification and development of the fat body, another mesodermally derived structure. The loss of SGPs and fat body development defects are also characteristics of loss-of-function mutants for the homeobox genes *tin* and *zfh-1* (Boyle et al., 1997; Moore et al., 1998_1; Broihier et al., 1998). This fact combined with the observed reduced expression of a GFP reporter driven by a *Six4*

enhancer in a *tin* mutant background (see section 1.14 and chapter 4) have led Clark et al. (2006) to believe that the roles of *tin* and *zfh-1* in fat body and SGP development are at least partly mediated by their regulation of *Six4*. Recent findings using ChIP arrays (Eileen Furlong, personal communication) as well as the results of this study (see chapter 4) are consistent with the regulation of *Six4* by Tin. Finally, $Six4^{289}$ homozygotes show severe muscle defects. Somatic muscles are seriously disrupted and disorganised in their association and attachment with some muscles being completely absent. This disruption seems to be caused by the of founder cells' inability to fuse. Founder cells are the distinct subset of myoblasts somatic muscle specification originates from. Finally, extensive disruption of the pharyngeal structure was observed in $Six4^{289}$ homozygotes coupled with the consistent disruption of certain muscle groups such as the segment border muscle (Ivan Clark, personal communication) (Figs 1.6 and 1.7).

The $Six4^{131}$ mutant also showed gonad coalescence defects of variable penetrance with phenotypes ranging from hypogonadism to the complete absence of gonads. The phenotype also includes scattering of non-coalesced germ cells even in the presence of a primitive gonad as well as testicular reduction in males and ovariole defects in females. $Six4^{131}$ homozygotes exhibit relatively mild muscle defects in comparison to $Six4^{289}$ mutants. No fat body defects have been reported thus far (Kirby et al., 2001; Clark et al., 2006).

The nature of these phenotypes identifies Six4 as a candidate for the regulation of numerous genes required for myoblast fusion, SGP-cell recognition and fat body specification and highlights potential similarities in the mechanisms facilitating these events. A candidate for Six4 regulation is the *Drosophila* muscle enhancer gene *ladybird* (*lbe*), which is required for the specification of a single muscle per embryonic hemisegment - the segment border muscle. This muscle is consistently affected in $Six4^{289}$ homozygotes and its disruption is a defining characteristic of the Six4 null phenotype. In wild type embryos Ladybird is expressed in cell patches called the pro-muscular cluster. In $Six4^{289}$/ $Six4^{289}$ these clusters are not present in most segments, a fact that hints towards the potential regulation of *lbe* by Six4 (Clark et al., 2006).

Another potential target of Six4 regulation is HMGCoA reductase (encoded by the *Drosophila* gene *columbus*). HMGCoA reductase mRNA was reported to be absent in the gonadal mesoderm of $Six4^{289}$/ $Six4^{289}$ embryos (Clark et al., 2006). HMGCoA reductase provides attractive cues to *Drosophila* germ cells, guiding them toward the

embryonic gonad (Santos and Lehmann, 2004) and its absence could be a possible explanation for the gonadal coalescence defects. Additionally the number of cells expressing SGP markers was found to be reduced in Six4 mutants, but these cells did appear to associate with germ cells. This data supports a role for Six4 in the specification or maintenance of SGP cell fate (Clark et al., 2006).

Although the exact phenotypic relationship between *Six4* and *SIX4/5* is still unclear, the defects in *Six4* mutant flies suggest that human *SIX5* might be a candidate for the muscle wasting and testicular atrophy phenotypes in DM1 and might have important functions in the development of mesodermally derived tissues. This theory is supported by the developmental effects caused by the disruption of *SIX5* homologues in other animal models. The fact that mouse knockouts of *Six4* (Ozaki et al., 2001) or *Six5* (Klesert et al., 2000; Sarkar et al., 2000*)* are viable, suggests that there might be extensive redundancy between the two genes. Similarly, mutations in the *unc-39* gene (previously named *ceh-35*) of *C. elegans*, a Six5 homologue, cause migration and differentiation defects in a subset of mesodermal and ectodermal cells, including muscles and neurons (Yanowitz et al., 2004). These defects include mesodermal specification and differentiation as well as neuronal migration and axon pathfinding defects in a manner reminiscent of the gonadal coalescence defects observed in *Drosophila*. This evidence points towards an involvement of SIX5 and its homologues in the processes of gonadogenesis and muscle specification through a potentially generic precursor cell guidance pathway.

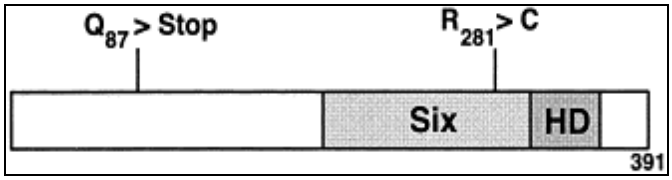


Fig.1.5 The predicted Six4 protein, including the SIX-(Six) and homeodomains (HD), with the molecular lesions identified for the two mutants (from Kirby et al., 2001)

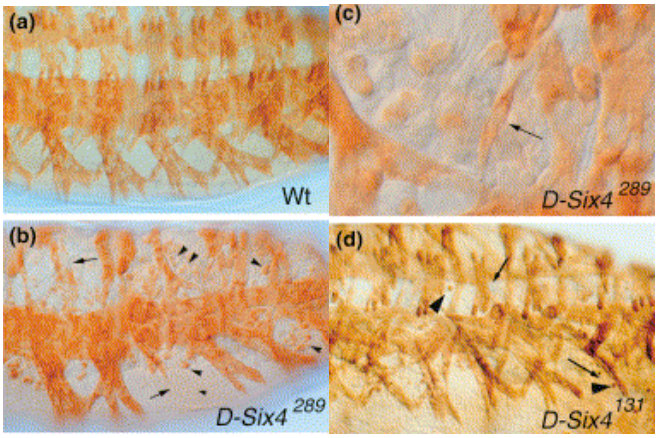


Fig. 1.6. Anti-myosin antibody staining of the abdominal musculature in wild type and *Six4*^{289}/ *Six4*^{289} embryos (designated here as D-Six4). Ventro-lateral views of abdomens of stage 16 embryos. (a) Wild-type embryo, showing the regular arrangement of syncytial myotubes. (b) A similar view of a homozygous *Six4*^{289} embryo, where muscles are highly disorganized; some are missing (arrows), and there are many unfused myocytes (arrowheads). (c) A higher power view of the region of (b), showing an apparently unfused (mononucleate) muscle founder cell (arrow). (d) A homozygous *Six4*^{131} embryo. Some muscle disruption and unfused myoblasts can be observed, even though such embryos are able to hatch (from Kirby et al., 2001)

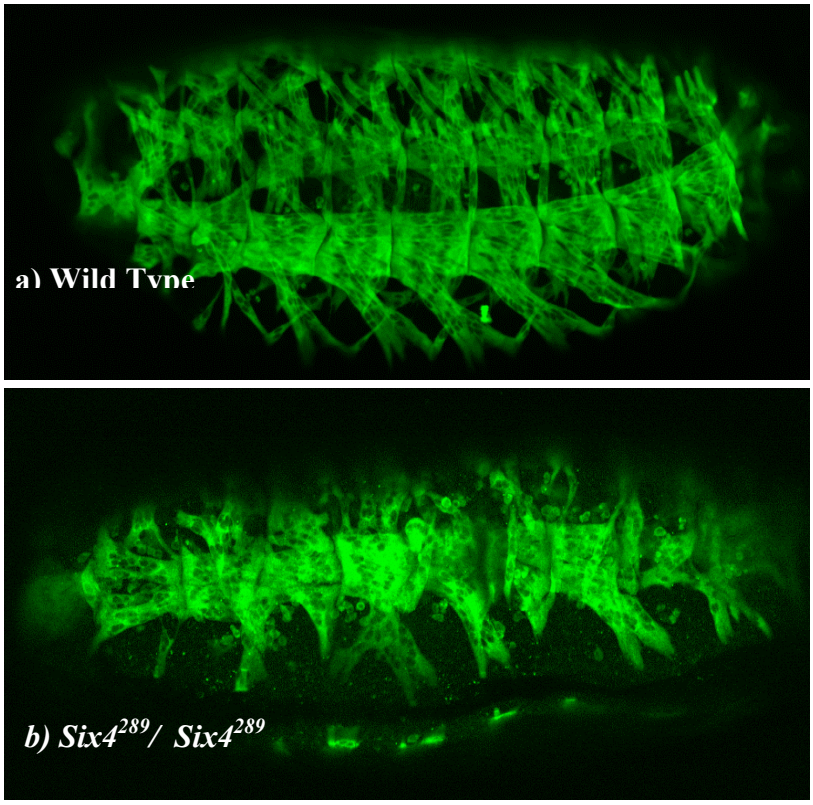


Fig. 1.7 Ventrolateral views of embryos stained with anti-myosin antibody a) Muscle arrangement in wild-type embryos b) Ventrolateral view of a *Six4*^{289}/*Six4*^{289} embryo (musculature is severely disrupted, segment border muscle is severely affected) (from Clark et al., 2006).

1.13 Six4 mesodermal expression pattern characterisation

The focus of this study is the elucidation of the role of Six4 in *Drosophila* development in general and the mesoderm in particular. As such, knowledge of its expression is important in identifying potential regulatory targets. The expression pattern of Six4 which is of particular interest to this study is described in greater detail in section 4.2. What follows is a brief description of the Six4 expression pattern.

The mesodermal expression of Six4 has best been described by Clark et al. (2006) through the use of a GFP reporter gene construct referred to as Six4-III-GFP (see chapter 4). Clark et al. (2006) identified an enhancer within the *Six4* third intron that activates GFP in a pattern corresponding closely to the mesodermal expression of Six4 RNA. These authors report that at stage 9, Six4-III-GFP is coexpressed with Mef2 in a broad mesodermal domain. Subsequently, by stage 10, GFP expression becomes restricted ventrally, with some minor protein presence in the dorsal region. The dorsal limit of Six4-III-GFP expression, after it becomes restricted, is identical to that reported for the Serpent (Srp) protein, a dorsolateral fat body cells marker (Abel et al., 1993; Baylies and Bate, 1996; Tapanes-Castillo and Baylies, 2004). ‘‘This anteroposterior modulation of D-Six4 expression resembles that of twi, raising the possibility that different levels of protein have different functional consequences’’ (Clark et al., 2006). At stage 10, inductive Dpp signalling from the dorsal ectoderm acts to maintain Tin expression, thereby driving the dorsal restriction of Tin (Staehling-Hampton and Hoffmann, 1994). The same reporter construct has subsequently been used by myself and the above observations have been verified. These findings are summarised in Fig. 1.8.

It has also been suggested by Clark et al. (2006) that the ventral restriction of Six4 may depend on an inhibitory effect of Dpp signalling. Consistent with this, misexpression of Dpp throughout the mesoderm reduces expression of Six4 RNA to a low level. Thus, it is suggested that ‘‘Dpp signalling acts to establish two, non-overlapping spatial domains of gene expression in the mesoderm: a dorsal domain expressing tin and a ventral and lateral domain in which D-Six4 is expressed’’ (Clark et al., 2006). Six4 is therefore a candidate for the counterpart of tin in patterning more ventral mesodermal fates.

As mentioned previously, Six4 is also required for fat body development, the other major organ arising from the non-dorsal mesoderm. Clark et al. (2006) reported a pronounced reduction in the number of fat body precursor cells in $Six4^{289}$/ $Six4^{289}$ embryos, thus identifying Six4 as a candidate for the specification and subsequent

maintenance of fat body precursor cells. Finally, Himeda et al. (2008) have recently demonstrated that the expression of Six4 in skeletal but not cardiac muscles is controlled by the myc-associated zinc finger protein MAZ.

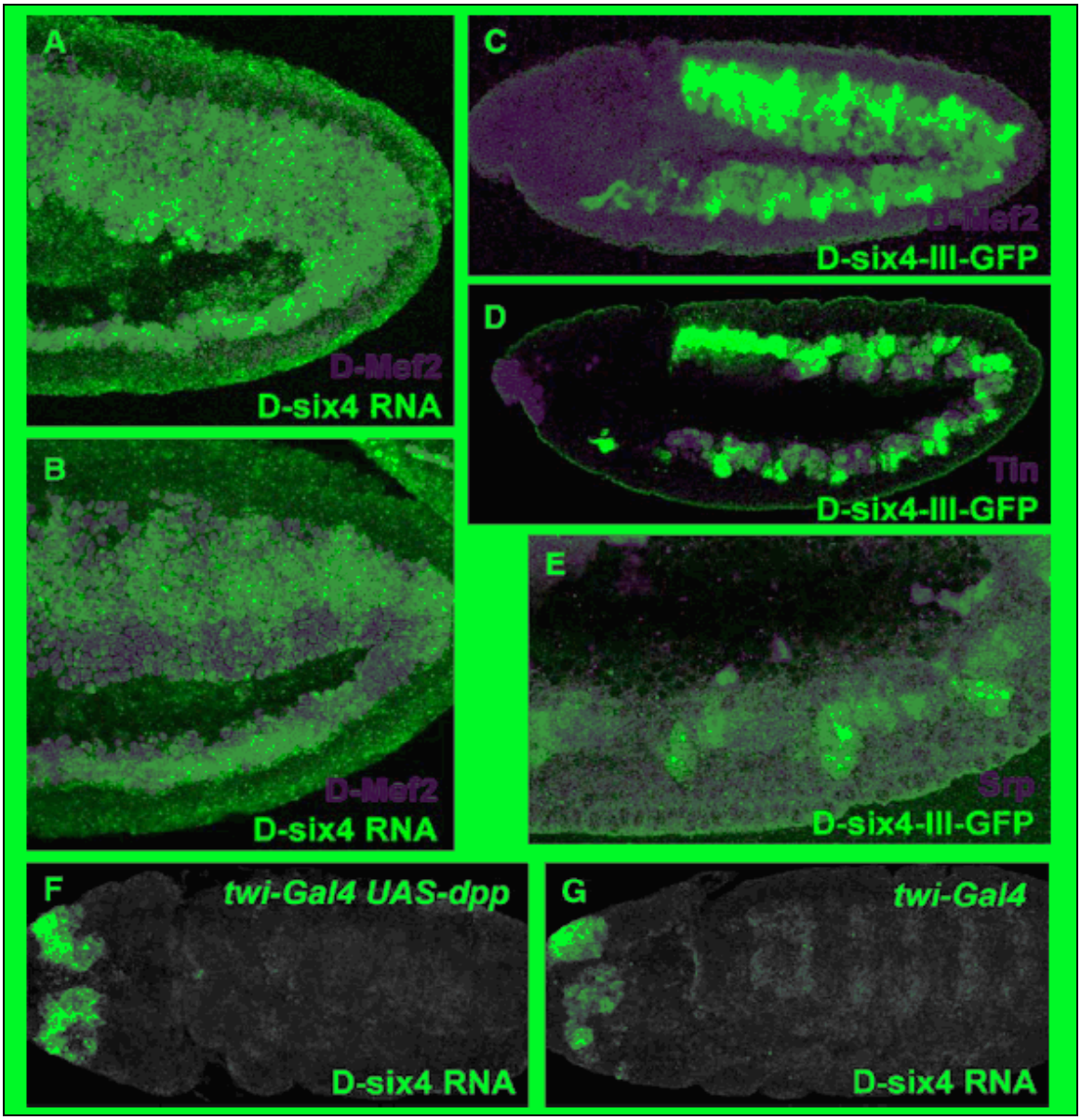


Fig. 1.8. Antibody staining of embryos showing restriction of *Six4* expression to the lateral and ventral mesoderm. (A, B) *Six4* mRNA expression in wild-type embryos, Mef2 (D-Mef2) is a panmesodermal marker. (A) Stage 9, *Six4* expression occurs throughout the mesoderm. (B), Stage 10, *Six4* RNA is lost from the dorsal mesoderm. (C–F) Embryos carrying a GFP reporter transgene driven by the *Six4* third intron enhancer (D-six4-III-GFP). (C) Stage 10. Relative to D-Mef2, stronger ventral/lateral GFP represents the restriction of *Six4* expression. Remaining dorsal expression represents GFP perdurance. (D) Stage 10, D-six4-III-GFP expression does not overlap with expression of Tin (magenta). (E) At stage 10 GFP is coexpressed with Srp, a fat body precursor specific marker. Six4 expression occurs in the dorsolateral eve domain but is more pronounced in the dorsolateral and ventral slp domain (seen here as the spaces between the Srp foci). (F) Picture of a twi-Gal4, UAS-Dpp embryo. Dpp misexpression results in reduction Six4 expression in the mesoderm Expression in ectodermally derived headremains unaffected. (From Clark et al., 2006)

1.14 Eyes absent (Eya), a Six4 co-factor

The developmental functions of Six4 appear at first glance to be dependent on the context specifying influence provided by other co-factors. As mentioned previously, members of the Eya class of phosphatases are known to be associated with SIX proteins. Specifically Six4 and sine oculis are known to interact with Eyes absent (Eya), also known as Clift (Cli).

The *Drosophila* Eya is a member of a protein family with known members in many metazoans (Jemc and Rebay, 2007). Eya proteins have been found in the cytoplasm and in the nucleus of cells in the embryo (Fougerousse et al., 2002), and SIX proteins are among the proteins that can transport Eya to the nucleus both *ex vivo* and *in vivo* (Grifone et al., 2005; Fan et al., 2000; Grifone et al., 2004; Ohto et al., 1999). Eya, like Six4 is required for somatic gonadal precursor development and its expression identifies somatic gonadal precursor cells (Clark et al., 2006; Boyle et al., 1997).

Eya proteins are characterised by a a phosphatase activity (Li et al., 2003; Rayapureddi et al., 2003) which is, however, not required for the transcriptional co-activator function performed in conjunction with sine oculis and has no effect in transcriptional output (Jemc and Rebay, 2007; Tootle et al., 2003). Eya proteins are also known participants in a MAPK/RTK signalling pathway in *Drosophila* (Hsiao et al., 2001), although this may not be the case in vertebrates. Vertebrate orthologues of these genes are not only expressed during eye formation where Optix acts upstream of Pax6 (Lagutin et al., 2003; Loosli et al., 1999) itself controlling the expression of Eya genes (Xu et al., 1997), but also during development of other organs including muscle (David et al., 2001; Heanue et al., 1999; Laclef et al., 2003; Sahly et al., 1999; Spitz et al., 1998), kidney (Xu et al., 2003), cranial placodes (Schlosser and Ahrens, 2004; Zou et al., 2004) and ear (Xu et al., 1999; Zheng et al., 2003). During kidney formation specifically, and possibly generally, *eya* genes lie upstream of *Six* and *Pax* genes, where they could cooperate with other *Hox* genes (Wellik et al., 2002). This observation, as well as the speculated necessity for Eya for Six4 function, reduces the likelihood of Eya being under Six4 regulation. It is however conceivable that SIX proteins can also act in an Eya independent manner and therefore the possibility of Eya being a target of Six4 cannot be ignored.

Eya1 is implicated in branchio-oto-renal syndrome, a dominantly inherited disorder characterized by hearing loss and branchial arch and renal anomalies in

humans (Ozaki et al., 2001). *Eya1*-deficient mice lack ears and kidneys, and heterozygous mutant mice show hearing loss and renal anomalies, as seen in human branchio-oto-renal syndrome. The presence of Six4 protein has also been found in acoustic ganglia and otic vesicles. These findings suggest that *Six4* could potentially be involved in the development of the ear in association with *Eya1*.

Despite evidence suggesting the contrary, Eya's phosphatase activity may be involved in transcriptional regulation. Phosphorylation and dephosphorylation are known to modulate transcription factor activity and is therefore consistent with the co-factor profile. Discoveries of dual-function proteins suggest that coupling an enzymatic activity directly to a transcription factor might be a way of regulating eukaryotic gene expression (Shi and Shi, 2004).

Scope of the thesis

With Six4 being a transcriptional regulator, its DNA binding specificity is integral to its function during *Drosophila* development. In spite of some suggestions (chapter 3) the Six4 recognition site remains largely unknown, as is its position within the mechanisms of gonadogenesis and muscle development. The identification of the Six4 binding specificity as well as that of the identity of potential downstream candidate genes will be essential in elucidating the developmental mechanisms underpinning *Drosophila* development. This is based on the premise that knowledge of *cis*-regulatory systems is essential in indicating both their internal workings and also the specific interconnections amongst them, i.e. the structure of the gene regulatory network. The aim of this work is to determine the binding specificity of Six4 and to identify potential Six4 regulation targets as well as to elucidate the regulation of *Six4* itself by other transcription factors through cis-regulatory elements in the immediate vicinity of the *Six4* locus.

Chapter 2 – Investigation of the DNA binding Specificity of Six4

2.1 Introduction

An analysis of the DNA binding specificity of the Six4 protein necessitates an in-depth understanding of the properties of homeodomain transcription factors in general and of SIX proteins in particular. As discussed in Section 1.9 the regulatory activity of Six4 is mediated by the sequence specific binding of its homeodomain to regulatory elements in the vicinity of its downstream target genes. As mentioned in previous sections the DNA binding specificity of Six4, like that of all the SIX proteins differs from that previously reported for most homeodomain transcription factors. Although little is know about the specificity of Six4, the binding site of its murine orthologues, Six4 and Six5, has been experimentally determined and was of importance in designing a Six4 binding specificity determination experiment.

2.2 Previously reported Six4/5 recognition sequences and targets

So far only a few genes have been identified as being regulation targets of murine Six4 and Six5. The ones that have been isolated suggest a diverse role for members of the Six4/5 subfamily in the regulation of various genes some of which are implicated in controlling muscle development.

The murine Six4 was originally discovered as the cell type-specific ARE (Na/K-ATPase_1 subunit gene regulatory element) binding factor AREC3. The murine AREC3 protein was found to be produced in the nucleus and cytoplasm of C2C12 myoblast cells and was shown to be augmented during muscle differentiation (Ohto et al., 1999; Kawakami et al., 1996; Suzuki-Yagawa et al., 1992). This is consistent with the involvement of Six4 in the muscle developmental pathway. More critically, ARE, Six4's original regulatory target is a member of a family of related P-type ATP-dependent ion transporter genes that includes the sarco-/endoplasmic reticulum Ca++-ATPase (Serca) genes *Atp2a1* and *Atp2a2*, for which altered expression levels have been reported in myotonic dystrophy (Damiani et al., 1996).

Additional Six4/Six5 regulatory targets have been identified through a screen for downstream targets of Six5 performed by Sato et al. (2002). This study revealed several candidate genes expressed in somites, skeletal muscle, brain and meninges, one of which was *Igfbp5*, encoding a component of IGF signalling. The overall expression level of Igfbp5 was found to be decreased in Six5-deficient mouse fibroblasts, and the response of human IGFBP5 to MyoD-induced muscle conversion

was altered in cells of DM1 patients (Sato et al., 2002). Other targets identified in this screen as well as their significance are fully discussed in section 3.7.

Six5 was also found to regulate Myogenin, a member of the MyoD family of proteins that is required for myoblast fusion *in vivo*. *MyoD* has been shown to be activated by Six1 and Six4 binding to the MEF3 motif present in the *Myogenin* promoter. However, due to the absence of Six4 in embryos at the time of *Myogenin* activation, the best candidates to control early activation of *Myogenin*, and thus early steps of myogenesis are Six1 and Six5 in conjunction with MEF2 and Myf5/MyoD proteins (Ruiz-Gomez et al., 2002; Spitz et al., 1998; Cheng et al., 1993).

The binding site of murine Six4 within the *ATP1A1* regulatory element (ARE) enhancer was identified through DNase I footprinting and methylation experiments as being GGTGTCAGGTTGC (conserved in human, mouse, horse and rat). The same experiments identified the a possible minimum sequence for binding as being GGNGNCNGGTTGC (Harris et al., 2000; Suzuki-Yagawa et al., 1992). The SIX domain and the homeodomain are both required for specific binding to GGTGTCAGGTTGC, although the homeodomain alone was shown to bind specifically to some other, unidentified, region of the ARE enhancer. This situation is reminiscent to that found in the Paired and POU classes of homeodomain proteins, in which the presence of two domains is required for specific DNA binding (Treisman et al., 1991). Murine Six2 and Six5 as well as *Drosophila* Six4 were also found to bind specifically to the Six4 binding site in the ARE (Kawakami et al., 1996). The sequence of this binding site is atypical because it doesn't contain a core tetranucleotide ATTA, something present in all previously reported homeodomain binding sites. The ATTA sequence interacts with an arginine at position five of the homeodomain (conserved in 95% of known homeodomains)(Gehring et al., 1994_2). However, SIX5 and SIX4 have a valine at this position and other members of the SIX family have serine or threonine. It is therefore likely that SIX homeodomains have a different binding specificity. Additionally, ''the amino acid at position 50 of the homeodomain normally recognises the two bases immediately 5' to the core sequence'' (Harris et al., 2000). Harris et al. (2000) used glutathione S-Transferase (GST)–SIX5 fusion proteins in gel retardation assays with short double stranded DNA fragments representing putative DNA binding sites, to investigate DNA binding targets of SIX5 and the functions of its two conserved domains in DNA binding specificity. It was shown in this study that SIX5 does not bind to a GGATTA consensus site present in the promoter of *DMPK* as it was previously suspected and

although a Six5 binding site was identified in the *DRD5* gene, no consensus sequence for the Six5 protein was found in that gene. Recombinant proteins containing the Six5 homeodomain were shown in the same study to form at least one specific complex with the ARE. A recombinant protein containing both the homeodomain and the SIX domains also formed a second specific complex with the ARE, assumed to be a dimer complex.

The identification of a consensus binding sequence for Six5 using random oligonucleotide selection (SELEX) (Rami Jarjour, personal communication) has also been reported. This sequence, which was reported as being CCGGTGTCTG, is highly similar to what was reported as being the ARE Six5 binding site. Additionally, Amphisix4/5, the amphioxus homologue of Six4 has been reported to bind to both the ARE sequence and the Six3 consensus binding site as well as the MEF3 myogenin binding site (Kozmik et al., 2007). These findings suggest that the binding specificities of orthologous Six4/5 subfamily members can potentially be similar to the point of being interchangeable.

It has also been independently determined through cell culture and transgenic studies that a DNA sequence previously known as the Trex site and which has subsequently been identified as being the Six5 binding site, and is important for *Muscle creatine kinase (MCK*) expression in skeletal and cardiac muscle is a target of the murine Six4 protein. Using gel shift assays and Six4-specific antisera, Himeda et al. (2004) demonstrated that Six4 binds to Trex in mouse skeletal myocytes and embryonic day 10 chick skeletal and cardiac muscle, while Six5 is the major TrexBF in adult mouse heart. ‘‘In co-transfection studies, Six4 transactivated the *MCK* enhancer (a 206-bp enhancer located from _1256 to _1050 bp upstream of the transcription start site) as well as muscle-specific regulatory regions of *Aldolase A, Myogenin* and *Cardiac troponin C* via Trex/MEF3 sites’’(Himeda et al., 2004). These results are consistent with Six4 being a key regulator of muscle gene expression in adult skeletal muscle and in developing striated muscle. The Trex/MEF3 composite sequence ([C/A]ACC[C/T]GA) has allowed for the identification of a novel putative SIX-binding site in six other muscle genes (Himeda et al., 2004). Gel shift experiments utilizing single-base-pair mutagenesis of the Trex site indicate that the sequence permitted for TrexBF binding is fairly degenerate, consisting of [C/A/T]A[C/T][C/T][C/T/G]GA[G/A/T] (C. L. Himeda and S. D. Hauschka, unpublished data referenced in Himeda et al., 2004).

A study by the same authors reported that although Six4 was the most likely Trex binding factor, it is conceivable that other factors such as CCHC-type zinc finger nucleic acid binding protein (CNBP), a protein involved in Type 2 myotonic dystrophy bind the Trex site in response to different physiological conditions (Himeda et al., 2004).

The exact DNA-binding requirements of SIX proteins are undetermined. However, ‘‘sequence database searches using the Trex/MEF3 composite sequence ([C/A]ACC[C/T]GA) revealed its presence in the regulatory regions of many muscle-specific genes. Two Trex/MEF3 sequences can be found in the mouse α-*Myosin heavy chain* Promoter, whereas one exists in the mouse β-*Myosin heavy chain* promoter and one in the human *Skeletal muscle α-actin* promoter. Four putative Trex/MEF3 sites are present in the rat *m1 Muscarinic acetylcholine receptor* (*AChR*) promoter, one is present in the rat *Neuronal AChR β* promoter, and two are found in the rat *Nicotinic AChR* promoter δ. Importantly, like the *MCK* Trex site. None of these putative binding sequences is a perfect match to the originally established MEF3 consensus’’(Himeda et al., 2004). All the sequences that are known to bind to members of the SIX4/5 subfamily are shown on table 2.1.

SIX5 Consensus	CAGA	C		A		C	C		G		G			
CNBP (reverse)		C		A		C	C		C		A		G	
MEF3 (reverse)		A		A		C	C		T		G		A	
TREX (permitted)		C/A/T		A		C/T	C/T		C/T/G		G		A	G/A/T
ARE	GC	A		A		C	C		T		G		A	C ACC
***MCK* TREX**		C		A		C	C		C		G		A	G

Table 2.1 Sequence comparisons between the SELEX-determined Six5 binding sequence (as defined by Rami Jarjour, personal communication), the CNBP (reverse consensus sequence), the MEF3 (reverse) consensus sequence (the element recognised by SIX proteins), the Trex permitted sequence (based on individual base pair mutation), the primary *ARE* enhancer sequence and the Trex sequence from the *MCK* enhancer. The underlined sequences have been shown to not be strictly required for SIX protein binding, although observations suggest that they are important for conferring binding specificity.

2.3 Determination of putative transcription factor binding sites

The primary goal of this study was the elucidation of the role of Six4 in *Drosophila* development through the identification of its downstream targets. As discussed previously, the developmental hardwiring inherent in all metazoans that allows for context specific expression of different proteins is implemented through the sequence specific DNA-protein interactions of the various transcription factors. If one is to decipher the function of a specific transcription factor and highlight its position within a developmental pathway, then knowledge of the specificity of these interactions becomes critical. Identification of a transcription factor's binding site (TFBS) is instrumental in placing that factor in a regulatory hierarchy. A TFBS, as well as the module containing it, constitutes the basic functional unit of transcriptional regulation.

Over the years a number of different approaches aimed towards establishing TF regulatory interactions have been utilised. This study will attempt to outline some of these approaches as well as to provide examples of regulatory information obtained through them. For a review of some of the methods outlined herein (as well as some others, with a strong focus on mammalian TFs) I would refer the reader to Elnitski et al. (2006). These approaches essentially fall into two categories: i) direct identification of the protein-DNA binding interactions, often in the absence of prior knowledge about the nature of the sequences identified by transcription factors or ii) experimental or theoretical determination of the DNA-binding specificity of a protein and subsequent identification of potential binding sequences within a genome for the identification of putative regulatory targets. This study will attempt to compare these methods in order to establish the preferred methodology for determining the DNA binding specificity of the Six4 protein given the existing circumstances and limitations. The methods outlined below are those most commonly used for protein specificity determination.

2.3.1 Footprinting

In the past many primary ligand binding sites have been determined by footprinting (Galas and Schmitz, 1978). ''Footprinting is essentially a protection assay in which the digestion of double-stranded DNA by a cleavage agent such as DNase I or hydroxyl radicals is locally inhibited by the binding of a ligand at specific binding sites within a DNA fragment''(Hampshire et al., 2007). Footprinting has in

the past been used to determine the binding specificities and kinetics of a number of known compounds such as actinomycin and distamycin A (Van Dyke et al., 1982). In such cases, details of the interaction with DNA can be elucidated further by crystallographic or NMR techniques. This technique however presupposes some knowledge of binding specificity and is unable to identify new transcription factor binding sites on a genomic scale. As such it is best used as an analytical tool for studying the properties of known ligand-DNA complexes rather than determining binding specificity *de novo*.

2.3.2 ChIP and ChIP-related approaches

Chromatin Immunoprecipitation (ChIP) follows in the footsteps of other approaches for elucidating ligand-DNA interactions *in vivo* such as *in vivo* footprinting, chemical and light-induced crosslinking and immunocytochemistry (Orlando, 2000). It involves the fixation (usually through formaldehyde cross-linking) and subsequent immunoprecipitation of protein-DNA complexes on chromatin. As such it offers the ability to detect any protein at its *in vivo* binding site directly. In particular, proteins that are not bound directly to DNA or that depend on other proteins for binding activity *in vivo* can be analysed with this method. ChIP was initially used in mammals for the identification of target genes of the HoxC8 and Oct4 proteins (Botquin et al., 1998; Tomotsune et al., 1993) as well as more recently for the identification of the binding sites of Polycomb, Trithorax and GAGA-factor proteins in the bithorax complex of *Drosophila* (Strutt et al., 1997; Orlando et al., 1997).

The main advantage ChIP has over analogous *in vitro* techniques is the ability to provide direct evidence that given regulatory proteins are associated 'in time and space' with specific genomic regions. In comparison, other methods outlined in this study (such as SELEX and recognition sequence modelling) provide indirect information about the potential 'occupancy' of a given site by a regulatory protein.

The concepts underpinning ChIP have also been applied in other techniques such as ChIP arrays and ChIP related approaches. These usually involve genome-wide screening procedures often combining the use of ChIP with that of DNA microarray analysis and were first used to identify binding sequences of the yeast transcriptional activator Gal4 (ChIP-on-chip or serial analysis of chromatin occupancy or SACO) (Ren et al., 2000; Impey et al., 2004). Recent applications of these techniques include the identification of regulatory targets of the mouse Stat3 protein, a part of the JAK-

STAT pathway, (Snyder et al., 2008), CREB, a transcription factor involved in cAMP signalling (Impey et al., 2004), and 3 members of the Hnf group of proteins (Hnf1, Hnf4, and Hnf6) (Odom et al., 2004).

However, in spite of the information volumes it can potentially generate, ChIP is a costly and cumbersome procedure that typically requires large numbers of cells, although recent advances claim to have reduced that necessity (Collas and Dahl, 2008; Dasgupta and Chellappan, 2007).

In addition to the techniques that can establish direct protein-DNA interactions, a number of techniques have been devised that can experimentally or theoretically determine the identity of a proteins preferred recognition sequence. This sequence can then be used to 'query' a known genome for instances of the desired motif.

2.3.3 SELEX

Unlike the methods described above, which provide direct evidence of ligand-DNA interactions, there are other methods that focus on determining the binding specificity of a ligand and then identify potential binding sites for that ligand in a genome. Putative TFBS identification using the data generated by these methods is carried out *in silico* (see chapter 3). One such method is Systematic Evolution of Ligands by EXponential enrichment (SELEX).

SELEX is also known by other names such as *in vitro* genetics, directed molecular evolution and cyclic amplification and selection of targets (Gold, 1995; Djorjevic, 2007). SELEX involves the use of large pools of oligonucleotides containing randomized sequences, the purification of ligand–nucleic acid complexes, and the amplification of nucleic acids contained in these complexes. Performing multiple cycles of this process generates a population of nucleic acids that are enriched in sequences exhibiting higher affinities for the ligand being investigated (these nucleic acids are called aptamers). Thus the purpose of the SELEX process is to minimize the number of background molecules and maximize the number of desired aptamer molecules.

Various types of SELEX have been used with differences primarily in the ligand-DNA complex purification process. Separation is achieved through the altered physical properties of ligand-bound nucleic acids (*e.g.*, reduced electrophoretic mobility), through ligand-specific affinity methods (*e.g.*, immunoprecipitation) or through the use of recombinant proteins (*e.g.* GST fusion proteins). Examples of

protein-DNA interactions characterised through SELEX include the identification of ssDNAs that bind to mammalian prion proteins (Bibby et al., 2008) the determination and of the consensus binding site for TFII-I Family Member BEN (Lazebnik et al., 2008), the identification of the binding sequence of the ESE-2 (Elf5) transcription factor (Choi and Sinha, 2006) and the determination of the DNA binding specificity of the Brn-3 proteins (Xiang et al., 1995).

Another variant utilising the same principles as SELEX is Restriction Endonuclease Protection, Selection and Amplification, or REPSA (Van Dyke et al., 2007). This process relies on the ability of bound ligands to inhibit an enzymatic cleavage process that would otherwise prevent unbound DNA from being amplified. Van Dyke et al. (2007) have used this approach to identify *de novo* the DNA binding specificities of at least 7 proteins with previously determined binding sites (also see Gopinath, 2007 for a very comprehensive list of aptamers isolated through various SELEX approaches).

Other types of SELEX may need to be employed when target protein requires the presence of the cell membrane (e.g., G-protein-coupled receptors, ion channels) or a co-receptor to fold properly. This is often the case when the ligand in question is a cell surface protein that can act as a therapeutic antagonist, agonist and/or diagnostic agent. In cases like this programming the SELEX experiment with purified, soluble protein target may be problematic. Shamah et al. (2008) present how this issue was addressed through the use of soluble membrane target ectodomains or complex mixtures such as membrane preparations or the surfaces of intact cells. This process is called complex target SELEX. These approaches are not related to this study since the role of Six4 in sequence recognition does not necessitate the presence of other factors. It is understood that the absence of Eya (see 1.14) does not alter the binding specificity of the Six4 protein (Ivan Clark, personal communication). The veracity of this claim has been confirmed by me (see below). Finally, Berezovski et al. (2006) have developed a method that is conceptually similar to SELEX. They have dubbed it non-SELEX and it does not involve the PCR amplification of aptamers between rounds but rather the partitioning of the initial aptamer pool through non-equilibrium capillary electrophoresis of equilibrium mixtures (NECEEM).

2.3.4 Protein binding microarrays

Another method that utilises the concepts applied in SELEX involves the use of protein binding microarrays (PBMs). This method involves a detectable (epitope tagged or directly fluorescent) protein of interest binding to a double-stranded DNA microarray. Binding affinity can be quantified through measurement of fluorescence intensity. The dsDNA array can be populated by synthetic sequences created in a randomised fashion reminiscent of SELEX (Linnell et al., 2004; Bulyk, 2006_1; Bulyk, 2006_2). This method has the same advantages as SELEX (and many of the same limitations) but allows for more direct quantification of binding affinity and can be considerably more rapid. Additionally it grants the experimenter complete control over the content of the tested aptamer collection. Accordingly, given its increased complexity, it can be more costly, and does not directly overcome the limitations of SELEX like discrepancies between observed and endogenous binding due to the absence of specificity-confering co-factors and the necessity for subsequent detection of *in vivo* binding sequences using *in silico* methods. This approach has successfully been used to examine the binding specificities of the Oct-1 transcription factor and the NF-κB p52 homodimer (Linnell et al., 2004).

2.3.5 Recognition sequence modelling

Finally, computational approaches exist that seek to determine a transcription factor's likely binding sequence based on its protein structure or *ab initio*. These methods rely on the premise that ligand-DNA interactions are governed by a ''Protein-DNA recognition code''. According to this notion binding sites of proteins can be predicted by knowing its amino-acid sequence and therefore attempting to predict its structure through homology to characterised proteins.

Although such a code was initially proposed (Seeman et al., 1976), it later became apparent that there is no simple deterministic recognition code. It was however demonstrated that certain amino-acids showed clear preferences in interacting with certain nucleotides (Matthews, 1988). Data-driven approaches that incorporate these preferences into probabilistic models have since been proposed and incorporated into algorithms that attempt to model relative interaction energies of DNA-recognition domains (SAMIE, Benos et al., 2001).

Similar approaches include the generation of binding site positional weight matrices (PWMs) by calculating the binding free energy differences for all possible single mutations within DNA recognition domain based on a 3D model of the protein-DNA complex. This method has been reported to accurately predict binding sequences for the yeast MAT-alpha2 homeodomain and GCN4 bZIP proteins (Liu and Bader, 2007)

However, these methods usually require accurate 3D structural models of transcription factor-DNA complexes or need further experimental validation if their findings are to be used in whole genome search for the predicted target sequence.

2.4 Rationale for the use of SELEX in the current study

SELEX was determined as being the desired approach for determining the DNA binding specificity of Six4 for a number of reasons. At the commencement of this study alternative methods were either unavailable (REPSA and non-SELEX) or were considered beyond the scope of this undertaking due to their high requirements in both resources and manpower (PBM and ChIP array) or were considered unsuitable for this purpose (footprinting and recognition sequence modelling). In the case of a ChIP array analysis in particular, the nature of the Six4 expression pattern necessitated that only the subset of cells that expressed Six4 be used in such an analysis. Selection of these cells, although in principle feasible through techniques such as fluorescent cell sorting (FACS) would complicate matters further. Another consideration was the requirement for a protein-specific antibody which had proven hard to obtain. The use of PBMs was also considered unsuitable given the fact that at the time of commencement of this study it was considered ‘‘impractical to create chips containing all DNA variants of 8-bp or longer’’ (Linnell et al., 2004). This limitation would prevent coverage of the predicted sequence space given the fact that previously reported Six5 binding sites were found to be up to 13 bps in length. Other approaches such as footprinting were deemed unsuitable for this pursuit due to the fact that they can’t really generate new information on a genomic scale.

Additionally, theoretical approaches such as recognition sequence modelling were hampered by the absence of data on related SIX protein homeodomains (which are sufficiently different from more common homeodomains to invalidate any potential comparative inferences) as well as by the questionable nature of the results generated through these approaches and the necessity for additional validation.

Finally, the DNA-binding specificity of the very closely related murine Six5 had previously been successfully determined through SELEX (Rami Jarjour., personal communication). Thus, SELEX was considered to be a straightforward and proven method that was not fraught with the difficulties and complications inherent in alternative approaches.

The benefits and limitations of SELEX are further discussed in the following sections. Also included is a more thorough analysis of the SELEX methodology.

2.5 SELEX Target Detection assay

As outlined above, ‘‘SELEX is an experimental procedure that allows extraction, from an initially random pool of oligonucleotides, of the oligomers with a desired binding affinity for a given molecular target’’ (Djorjevic, 2007). This section will provide a more detailed view of the process of SELEX.

SELEX was developed as a sensitive and rapid method of determining the sequence specificity of DNA binding proteins (Pollock and Treisman, 1990). It allows for recovery of targets using protein present is crude cell extracts or purified samples. The procedure is used to infer the strongest binders for a given DNA or RNA binding protein, and the highest affinity binding sequences isolated through SELEX can have numerous research, diagnostic and therapeutic applications (Djordjevic, 2007). Authors alternatively refer to SELEX as SAAB for selected and amplified binding sites (Blackwell et al., 1990), CASTing for cyclic amplification and selection of targets (Wright et al., 1991), or simply in vitro selection (Oliphant et al., 1989; Ellington and Szostak, 1990).

SELEX relies on a conceptually straightforward method. A starting oligonucleotide pool is generated in a standard DNA-oligonucleotide synthesizer. This oligonucleotide will contain a completely random base-sequence which is flanked by defined primer binding sites. As such the range of sequences covered in the initial oligo pool depends on the parameters of synthesis but usually ranges between 10^{14}-10^{15}. The immense complexity of the generated pool justifies the assumption that it contains a few molecules with the correct sequence in the case of DNA binding proteins or the correct receptor structure or with tertiary structures which lead to catalytic activity in the case of RNAs. These aptamers, as they are called are then selected through methods such as affinity chromatography or filter binding. Because a pool of such high complexity can be expected to contain only a

very small fraction of functional molecules, several purification steps are usually required. Therefore, the best binding molecules are in principle amplified by the polymerase chain reaction (PCR) or in a transcription-based step using known primers that anneal to sequences included in contents of the oligonucleotide pool. In this way, iterative cycles of selection can be carried out. Successive selection and amplification cycles result in an exponential increase in the abundance of specifically binding sequences, until they dominate the population (Djordjevic, 2007). Various considerations need to be addressed in the design of a SELEX experiment. These include the initial stoichiometry of the reaction and the selection dynamics that directly influence round to round distributions of nucleic acid fractions in the reaction (these fractions are binding and non-binding oligos i.e. aptamers and non-aptamers). A mathematical analysis of the process has been performed by Levine and Nilsen-Hamilton (2007) but optimisation of the reaction beyond the determination of the initial parameters is very much a trial-and-error process.

SELEX was used in this study to highlight the role of Six4 in *Drosophila* development by initially determining its DNA binding specificity *in vitro*.

2.6 Experimental Aims

In light of what is already known about Six4 the aim of this project was to establish and test the sequence requirements for DNA binding to Six4 by an *in vitro* selection procedure using random oligonucleotides to isolate sequences with an affinity for Six4. The resulting binding sites were used to screen the *Drosophila* genome for potential binding sites which can then be tested for binding to Six4 in an attempt to elucidate the regulatory role of Six4 (as well as its human orthologue SIX5) in *Drosophila* development.

2.7 Experimental design

As discussed earlier, SELEX requires the separation of protein-DNA complexes after each selection round. In this study this was achieved through the use of a GST fusion protein. The protein in question is a recombinant protein created using the p-GEX system that incorporates the DNA binding domain of Six4 coupled with a glutathione S-transferase that facilitates binding to glutathione sepharose beads and allows for isolation of DNA-protein complexes. The SD and HD of Six4 were amplified from genomic DNA (for primer sequences see chapter 5) using the topology

outlined in Seo et al. (1999) and expressed as part of a glutathione S-transferase recombinant protein. Protein extracts were co-incubated with a pool of oligonucleotides that initially contained a 26bp random core (equal chance for any nucleotide at any position), that act as a source of potential binding sites, flanked by two known 25bp sequences, that anneal to known primers (oligos primer R and primer F, for sequences see section 5.1.2.6) and allow for PCR amplification. DNA:protein complexes were then isolated through binding to glutathione sepharose beads (Fig. 2.1). This process constitutes a selection round. Selected sequences were then amplified by PCR and used in subsequent selection rounds, thus creating an oligonucleotide pool enriched for sequences that show affinity for the recombinant protein (Fig. 2.1). The SELEX rounds were repeated several (3-5) times, and some of the oligonucleotides selected in the final round of the experiment were sequenced and their sequences aligned in order to define the Six4 binding site consensus.

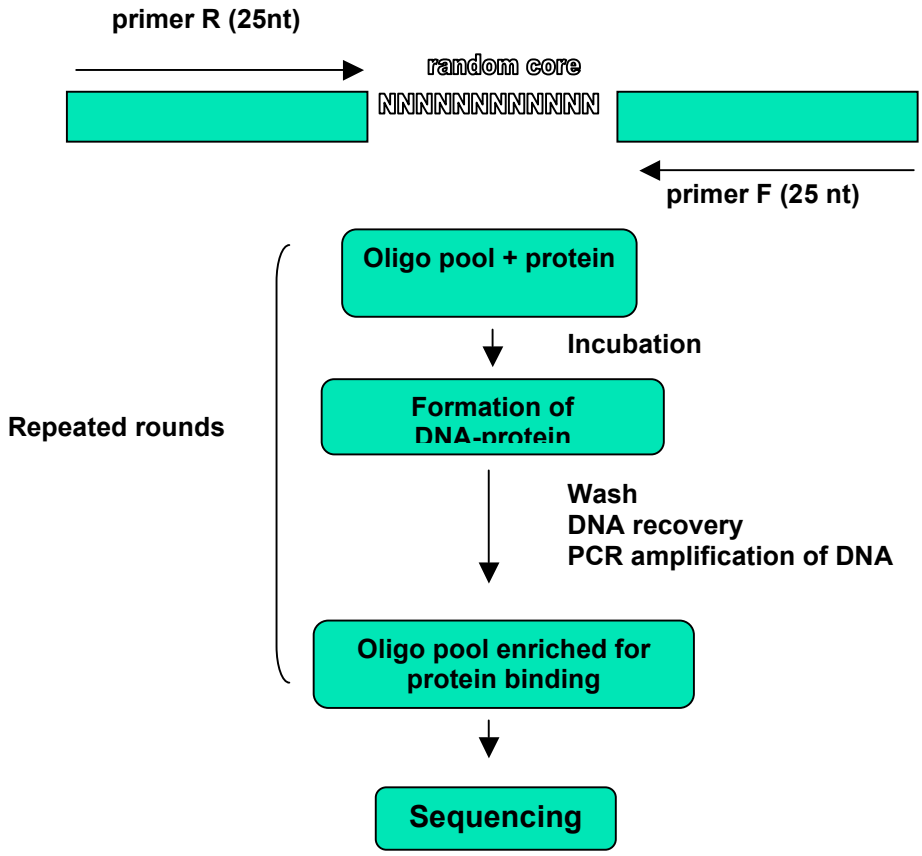


Fig. 2.1. A scheme of the standard SELEX procedure. The starting pool of sequences consists of random oligonucleotides. The oligonucleotide pool is then gradually enriched for the high affinity binders, by repeated rounds of target molecule binding, selection and amplification. After a certain number of rounds is performed, some of the oligos selected in the last round of the experiment are sequenced

2.8 GST-SD+HD Recombinant Protein expression and purification

The recombinant protein used in this study was created using the GST gene fusion system which allows for the expression, purification and detection of glutathione S-transferase fusion proteins using *E.coli* (Ausubel et al., 1996). This expression vector system enabled the protein glutathione S-transferase (GST) (26 kDa) to be produced in fusion with the protein of interest so as to bind with high affinity to a glutathione sepharose beads.

The sequence encoding the Six4 homeo- and SIX domains was determined through sequence alignment based on the topology previously established by Seo et al. (1999) (Fig. 2.2). According to these authors the SIX and homeodomains of Six4 were determined to be located in positions 176-295 and 296-351 respectively (Flybase protein ID is FBpp0077932). The sequence spanning the SD and HD was amplified from cDNA using primers containing the BamHI and EcoRI restriction sites (for primer sequences see 5.1.2.6) and cloned in frame in the p-GEX(2T)^{TM} expression plasmid. This was determined as being the largest span of Six4 that could be incorporated into a GST fusion system whilst still resulting in a soluble, and therefore functional protein (Feng Li, personal communication), that could accurately emulate the Six4 DNA binding specificity when expressed as part of a GST fusion protein. This claim was made on the basis of the protein showing specific binding to an oligonucleotide containing the Six5 binding sequence found in the ARE enhancer (AREo, see section 5.1.2.6**)**. This fact was independently demonstrated by Feng Li and Ivan Clark (personal communication) and was subsequently repeated by myself (Fig. 2.7).

A recombinant plasmid containing the SIX and homeodomain encoding sequences of Six4 and a glutathione S-Transferase N terminus was then generated using the p-GEX(2T)^{TM} GST gene fusion system (Kaelin et al., 1992)(Fig. 2.3) and was then transformed in the XL1 blue expression line of *E.coli*. (For cloning and bacterial transformation protocols see chapter 5). Resulting colonies were cultured overnight in LB broth. Plasmid DNA was extracted from these cultures and subsequently subjected to restriction analysis (Fig. 2.4) using the BamHI and EcoRI restriction enzymes so as to cause excision of the inserted SIX- and homeodomain encoding sequence. Clones showing successful excision of the insert from the expression vector (lanes 2-13 in Fig. 2.4) were subsequently sequenced to confirm the successful insertion of non-mutated Six4 SD and HD sequence. This was confirmed to be the case in 6 out of 12

clones. The plasmid containing the SIX and homeodomain encoding sequence (resulting from clone 4) was transformed into the BL21 expressing line of *E.coli* through electroporation (see chapter 5). The BL21 strain is defective in OmpT and Lon protease production and will therefore increase soluble protein production (see below).

Cells containing the GST-SD+HD expressing plasmid were cultured overnight (for expression conditions see below) and protein expression was subsequently induced through the addition of isopropyl-beta-D-thiogalactopyranoside (IPTG). The soluble fraction of whole cell extracts obtained through sonication was found through SDS-gel electrophoresis, to contain a protein of a size corresponding to that expected of GST-SD+HD (46 KDa, Fig. 2.5). An additional induced band can be seen at circa 30 kDa in figures 2.5 and 2.6.*. A likely explanation for the presense of this band is that the fracture of the recombinant protein as a result of over-sonication during protein extraction.

```
1   MFDKNLDGNNLSVSIGGDLDSTSSGGTSSDHSAVHQDNLSSPMAYGSLFL
51  PNAGYRGNLSCKTVLQLDKFAPYEGVEKDHLLERRFQDITNDYDKSPPPT
101ASTTPTHYPSLNSIIFENGSSGNLGDLNGNTKTDLCAGLQRSGGGLGGNA
151GSGGHLISNLTAAHNMSAVSSFPIDAKMLQFSTDQIQCMCEALQQKGDIE
201KLTTFLCSLPPSEFFKTNESVLRARAMVAYNLGQFHELYNLLETHCFSIK
251YHVDLQNLWFKAHYKEAEKVRGRPLGAVDKYRLRKKYPLPKTIWDGEETV
301YCFKEKSRNALKDCYLTNRYPTPDEKKTLAKKTGLTLTQVSNWFKNRRQR
351DRTPQQRPDIMSVLPVGQLDGNGFPRMFNAPSYYPETIFNGQ
```

Fig. 2.2 Primary amino-acid sequence of Six4. Amino-acids in **pink** and **blue** represent the SIX- and homeodomains respectively and are included in the recombinant GST fusion protein. The above topology is presented as reported by Seo et al. (1999).

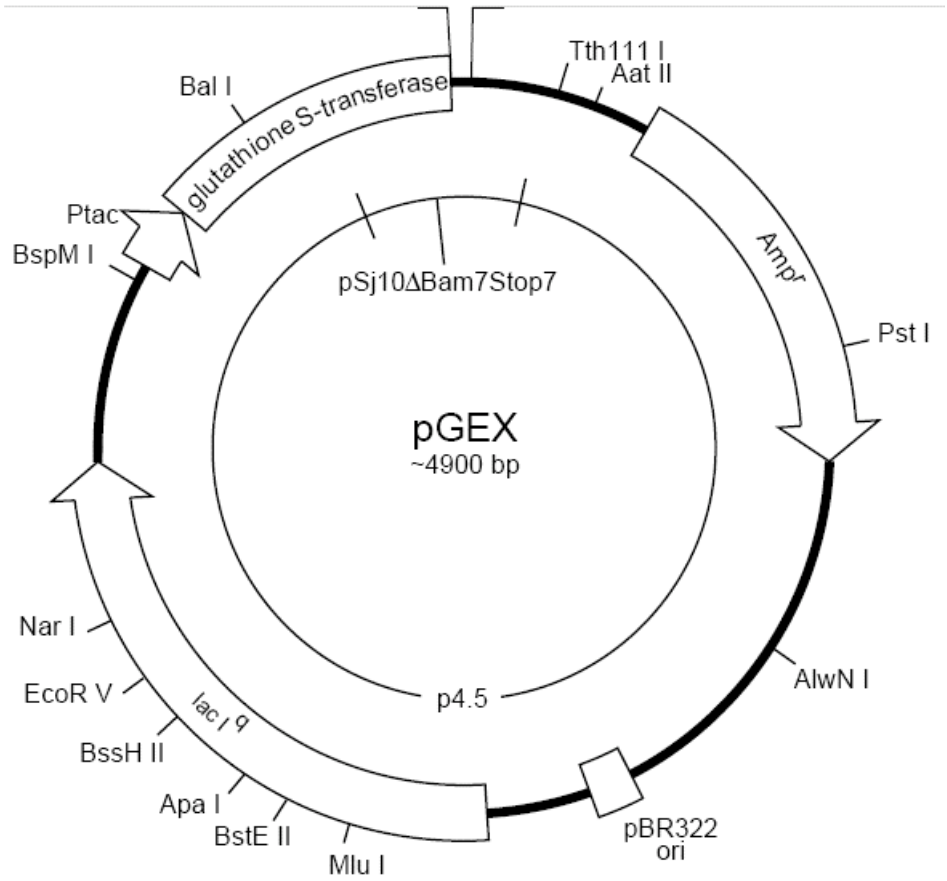


Fig. 2.3 The pGEX cloning vector was used for expression of the Six4-GST recombinant protein. The glutathione S-transferase (GST) Gene Fusion System was used for the expression, purification, and detection of fusion proteins produced in *Eschericia coli*. This resulted in inducible, high-level expression of the Six4 homeo- and SIX domains as fusions with *Schistosoma japonicum* GST (Smith and Johnson, 1988). Expression in *E.coli* yielded fusion proteins with the GST moiety at the amino terminus and the protein of interest at the carboxyl terminus. This resulted in a Mr 46 000 protein that was expressed in *E.coli* with DNA-binding activity. Fusion proteins possess the complete amino acid sequence of GST and therefore demonstrate GST enzymatic activity and can undergo dimerization similar to that observed *in vivo* and can thus bind to glutathione coated sepharose beads. Originally the ranges of the homeo- and SIX domains were those first described by Seo et al. (1999).

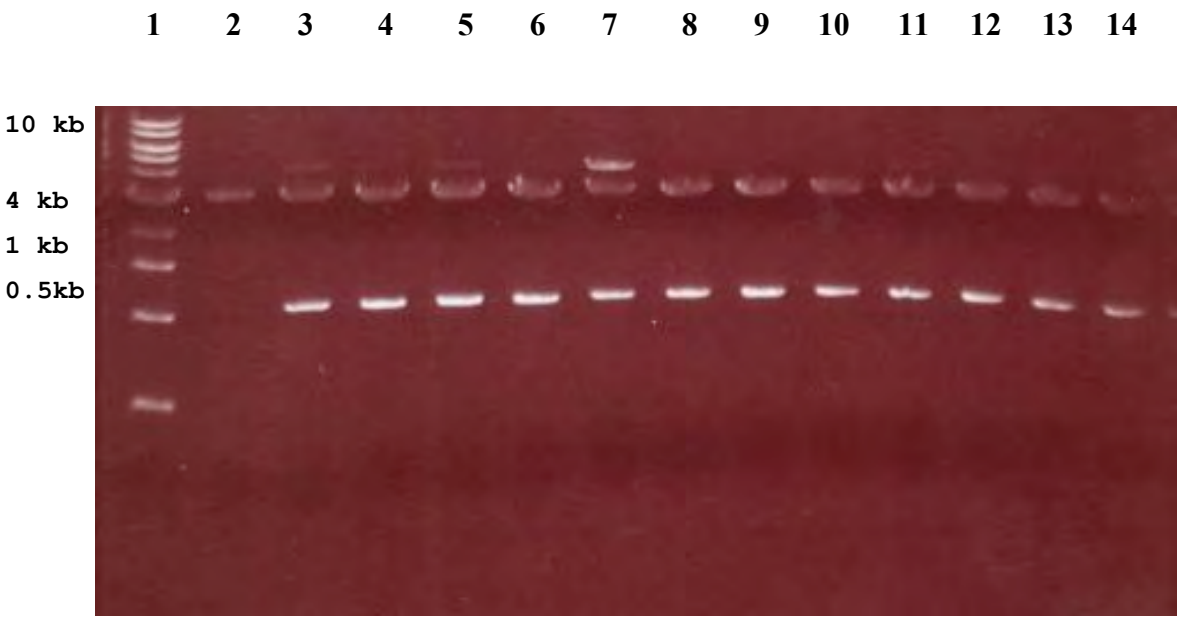


Fig. 2.4 Agarose gel electrophoresis of a restriction analysis of the GST-SD+HD plasmid. Plasmid DNA recovered from single colonies transformed with the Six4 homeo- and SIX domain encoding sequence ligated into the pGEX cloning vector. Plasmid DNA was digested with BamHI and EcoRI restriction enzymes (lanes 3-13) the restriction sites of which can be found on the extreme flanks of the inserted sequence. Restriction results in linearization of the pGEX vector (4.5 kb band) and the exclusion of the cloned insert (~500bp fragment). Lane 1 contains size markers ranging from 10 kb to 0.5 kb (NEB 1kb marker ladder). Lane 2 contains pGEX cloning vector digested withBamHI and EcoRI.

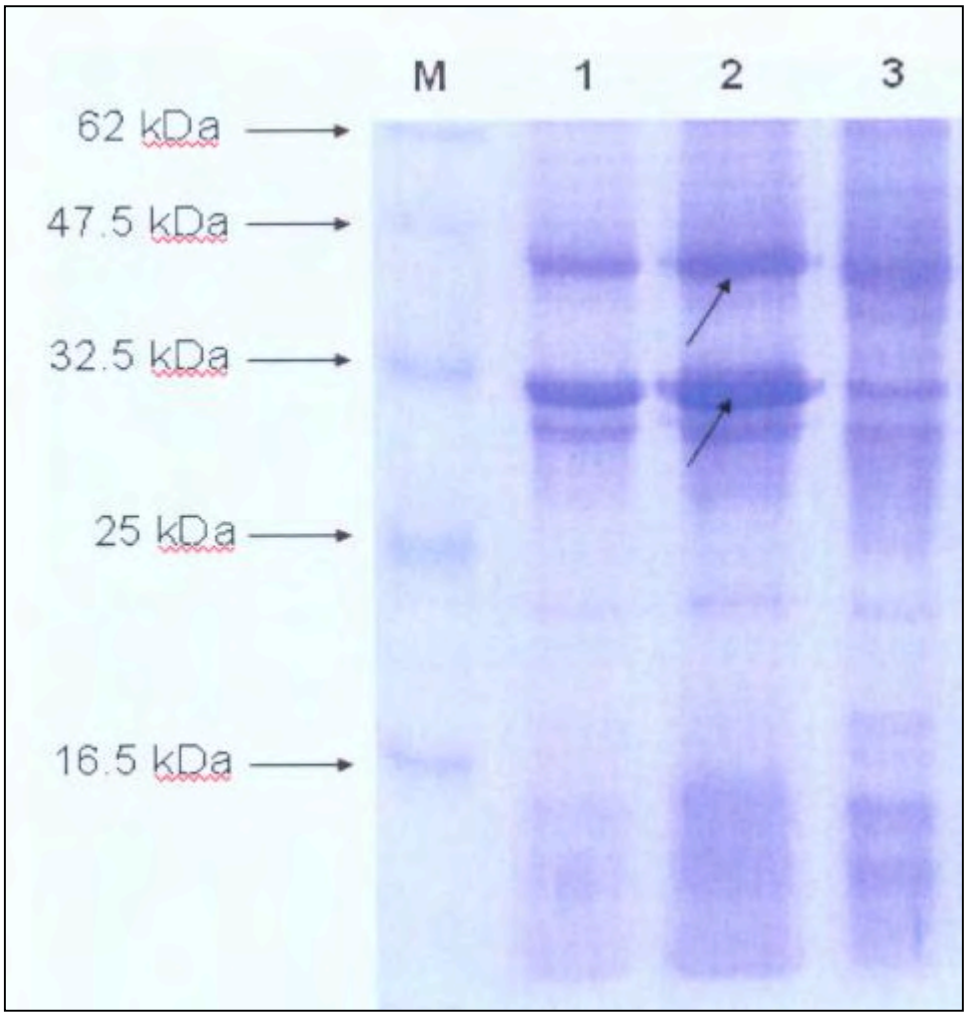


Fig. 2.5 SDS-PAGE analysis of soluble cell extracts from cultures induced by IPTG. Lane M , Perfect Protein Marker Ladder; Lanes 1-2 WCE of cells transformed with the GST-SD+HD plasmid and grown overnight until they reached ODs of ~0.7 (lane 1) and ~1 (lane 2) and then grown in medium supplemented with 0.1mM IPTG (equal volumes loaded, equal cell extract mass loaded). Lane 3 WCE of cells grown in the absence of IPTG. Arrows indicate the induced bands. The upper band corresponds to the recombinant GST-SD+HD protein. The lower band is likely to correspond to the GST moiety (roughly 26 kDa) bound to at least part of the SD+HD. This is likely a product of over-sonication

Due to the presence of the SIX domain and homeodomains the recombinant protein was expected to accurately emulate the DNA binding specificity of the original Six4 protein and therefore allow for its study *in vitro*. As is discussed in sections 1.10 and 1.15 SIX proteins are known to synergise with co-factors. However, the absence of these co-factors has not caused alteration in the binding specificities of any of the previously reported SIX proteins (Rami Jarjour, personal communication; Sato et al., 2002).

Expression conditions and parameters were carefully regulated to ensure a sufficient yield of protein for subsequent applications whilst preventing excessive expression from causing the resulting proteins to form aggregates that can be trapped in precipitates (inclusion bodies). The formation of inclusion bodies is a common side-effect of recombinant protein expression. Scopes (2001) reviews the methodology of protein purification and addresses inclusion body formation. In spite of their order of appearance in this thesis, the optimisation of soluble protein expression was in fact performed after initial SELEX experiments highlighted the necessity for highly concentrated soluble protein extracts.

The creation of insoluble, and therefore unusable, aggregates of protein depends on the concentration of protein present in the induced culture. Factors controlling protein concentration include the concentration (expressed as OD) of protein expressing cells, the concentration of IPTG (and its regulatory effect on protein expression), the time of incubation post-induction (longer induction times allow for greater accumulation of protein) and temperature (temperatures below 37^0 typically result in higher yields of soluble protein). Optimisation of these parameters is often done in a trial-and-error fashion with soluble protein yield being the criterion based on which different conditions are assessed. For the purposes of this study protein, concentration was measured through a Coomassie dye-based (Bradford) protein assay (see chapter 5). The absorbance at 595 nm of the soluble fraction of whole cell extracts (WCEs) coincubated with 30-fold excess of Bradford reagent was used as a means of assaying soluble protein yield. Table 2.2 summarises the protein yield under different sets of conditions. What follows is a description of the different condition sets used in protein expression optimisation.

Cultures where grown overnight until they acquired optical densities (OD) of 0.19 -1.3 (0.19, 0.25, 0.5 and 1.3). Recombinant protein expression was induced using 0.05 or 0.1 mM IPTG at 30^o and culture samples were obtained 2, 3.5 and 5 hours post

induction. Induced and uninduced control cells were harvested and fragmented through sonication. GST-SD+HD recombinant protein was obtained through purification of soluble cell extracts on glutathione sepharose beads and subsequent protein elution with glutathione (see section 5.6.1). Relative concentrations were determined both through arbitrary comparison of SDS gel band intensities of WCEs and a Bradford assay of purified protein concentration (Bradford, 1976) (Table 2.2, Figs. 2.6.*). Based on the findings summarised on table 2.2 the optimal protein expression condition were determined to be those detailed in chapter 5 (OD pre induction is 1.3, concentration of IPTG is 0.05mM and post induction incubation time is 3.5) since the purified protein sample obtained under those conditions was found to have the highest absorbance when reacting with the Bradford reagent (0.122).

No direct conclusions were drawn based on the influence these parameters exerted on soluble protein yield as the observed trends in protein expression are products of a combination of factors. A higher initial OD may reduce protein production by causing the cultures to enter lag phase prematurely. This factor is in turn linked to post induction incubation time. Observations suggest that longer incubation times (5 hours) reduce soluble protein yields, possibly through the formation of aggregates. Inversely, incubation times of 2 hours result in relatively high soluble protein yields. It is indeed conceivable that cultures that were incubated for only 2 hours may contain the highest concentrations of soluble expressed proteins in proportion to the concentration of cells in the culture. Similarly, increases in the concentration of IPTG in the culture generally increase the yield of soluble protein. However, since all these factors essentially control the same variable (the concentration of soluble fusion protein) through different means, they cannot be addressed individually. This optimisation analysis is of limited usefulness in determining the exact effect these parameters have in soluble protein expression. In spite of this shortcoming however it is sufficient to establish the desired set of conditions (out of those tested) for optimal protein expression and as such serves the purposes of this study. A more detailed analysis that involves the regular measurement of the fraction of GST-fusion protein that exists in the soluble and insoluble states in response to careful modulation of one of the aforementioned parameters (and perhaps including induction temperature) may establish the dynamics of protein aggregation and allow for a more informed decision to be made in order to determine the best expression conditions.

Optical Density of culture prior to induction	Concentration of IPTG in mM	Post induction incubation time in hours	Absorbance at 595nm
0.5	**0.05**	**3.5**	**0.063**
0.5	**0.1**	**3.5**	**0.085**
0.5	**0.05**	**5**	**0.012**
0.5	**0.1**	**5**	**0.042**
0.5	**0.05**	**2**	**0.062**
0.5	**0.1**	**2**	**0.085**
1.3	**0.05**	**2**	**0.087**
1.3	**0.1**	**2**	**0.051**
1.3	**0.05**	**3.5**	**0.122**
1.3	**0.1**	**3.5**	**0.114**
1.3	**0.05**	**5**	**0.031**
1.3	**0.1**	**5**	**0.045**
0.25	**0.05**	**2**	**0.04**
0.25	**0.1**	**2**	**0.059**
0.25	**0.05**	**5**	**0.016**
0.25	**0.1**	**5**	**0.019**
0.25	**0.05**	**2**	**0.035**
0.25	**0.1**	**2**	**0.061**
0.19	**0.05**	**3.5**	**0.035**
0.19	**0.1**	**3.5**	**0.067**
0.19	**0.05**	**3.5**	**0.017**
0.19	**0.1**	**3.5**	**0.036**
0.19	**0.05**	**5**	**0.032**
0.19	**0.1**	**5**	**0.04**

Table 2.2 Table of the absorbance(at 595 nm) of purified protein extracts (0.05 ml) obtained under a range of expression conditions and incubated with 30-fold volume excess of Bradford reagent (1.5 ml). The highest observed absorbance is underlined (9th from top).

1 2 3 4 5 6 7 8

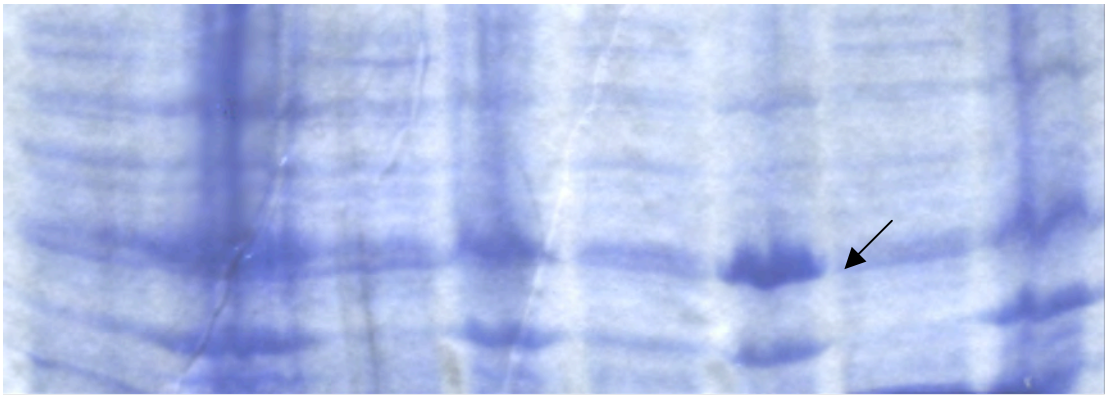


Fig. 2.6.1 OD prior to induction: 0.5, Lane 2: 0.05mM IPTG-3.5h, lane 4: 0.1 mM IPTG-3.5 h, lane 6: 0.05mM IPTG-5h, lane 8: 0.1mM IPTG-5h

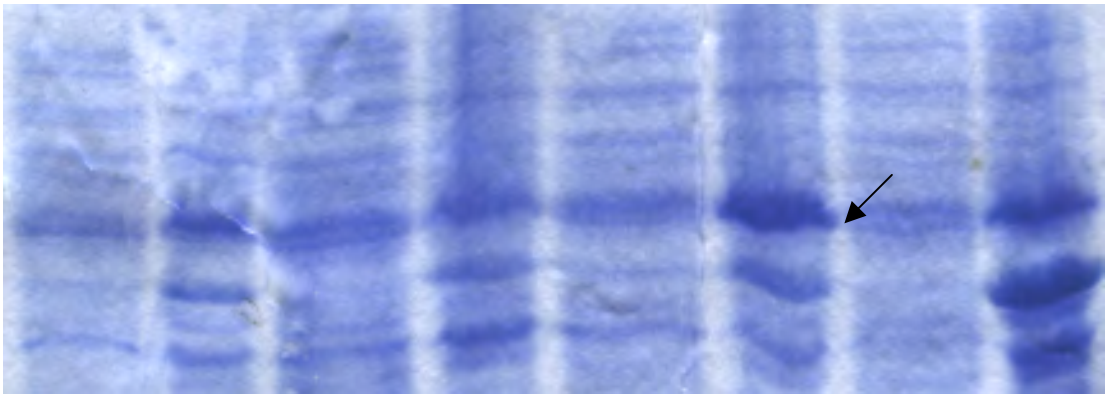


Fig. 2.6.2 OD prior to induction: 1.3, Lane 2: 0.05mM IPTG-2h, lane 4: 0.1 mM IPTG-2h, lane 6: 0.05mM IPTG-3.5h, lane 8: 0.1mM IPTG-3.5h

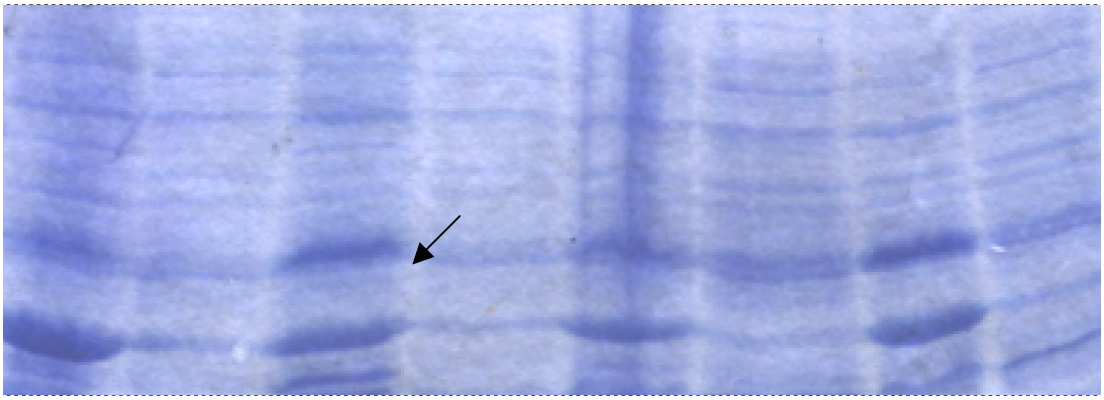


Fig. 2.6.3 OD prior to induction: 0.25, Lane 1: 0.05mM IPTG-2h, lane 3: 0.1 mM IPTG-2h, OD prior to induction: 0.19, lane 5: 0.05mM IPTG-3.5h, lane 7: 0.1mM IPTG-3.5h

Fig. 2.6 SDS-PAGE analysis of soluble fraction of WCE from cultures induced by IPTG (equal cell extract mass loaded)

Odd numbered lanes represent uninduced control samples grown under the conditions applying to the even numbered lane that succeeds them i.e. lane 5 in any given figure represents the uninduced culture grown under the conditions that apply to lane 6. Black arrows indicate induced expression.

1 2 3 4 5 6 7 8

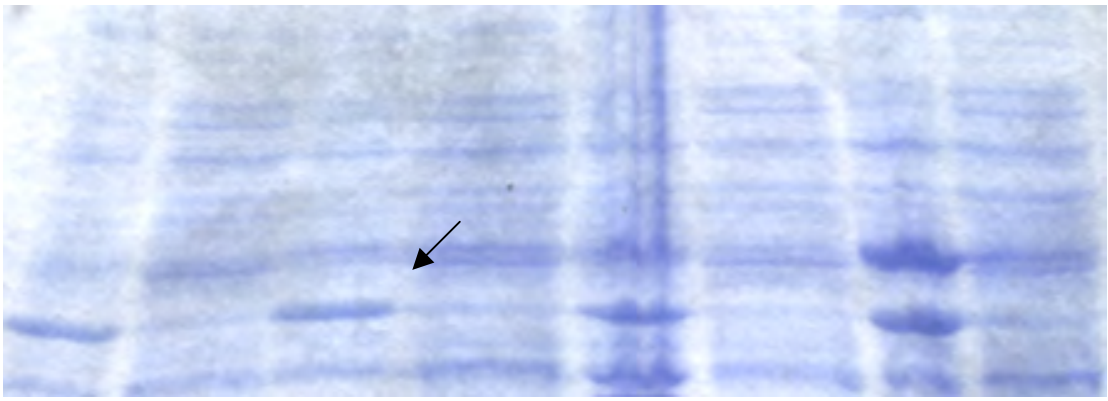


Fig. 2.6.4 OD prior to induction: 0.25, Lane 1: 0.05mM IPTG-5h, lane 3: 0.1 mM IPTG-5h, lane 5: 0.05mM IPTG-2h, lane 7: 0.1mM IPTG-2h

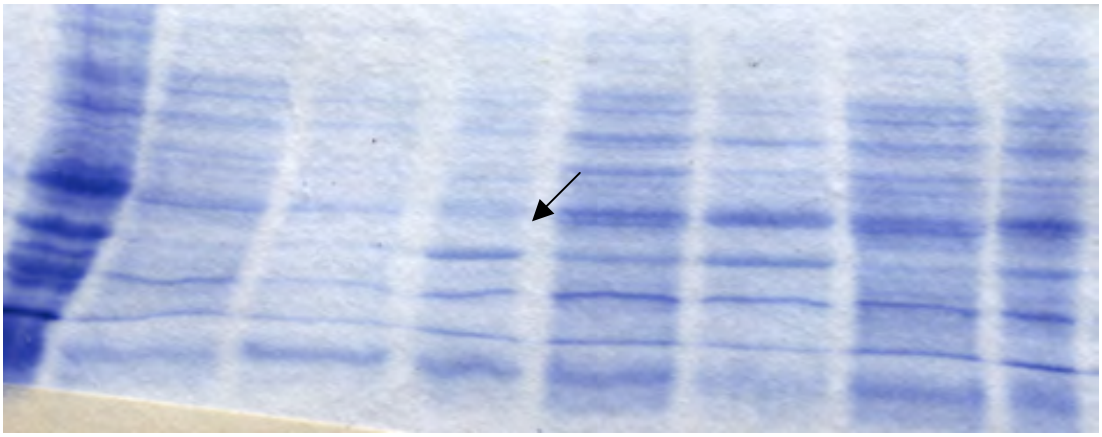


Fig. 2.6.5 OD prior to induction: 0.19, Lane 2: 0.05mM IPTG-3.5h, lane 4: 0.1 mM IPTG-3.5h, lane 6: 0.05mM IPTG-5h, lane 8: 0.1mM IPTG-5h

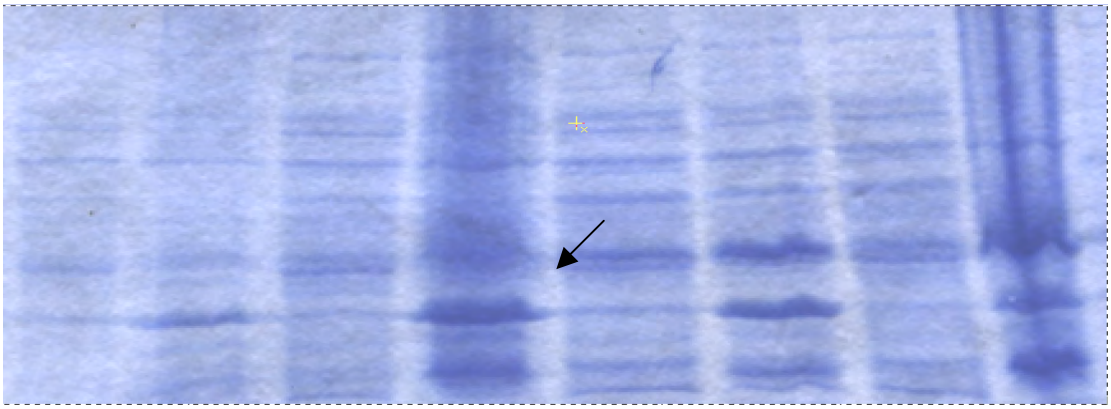


Fig. 2.6.6 OD prior to induction: 1.3, Lane 2: 0.05mM IPTG-5h, lane 4: 0.1 mM IPTG-5h, OD prior to induction: 0.5, lane 6: 0.05mM IPTG-2h, lane 8: 0.1mM IPTG-2h

Fig. 2.6 SDS-PAGE analysis of soluble fraction of WCE from cultures induced by IPTG (equal cell extract mass loaded)

Odd numbered lanes represent uninduced control samples grown under the conditions applying to the even numbered lane that succeeds them i.e. lane 5 in any given figure represents the uninduced culture grown under the conditions that apply to lane 6. Black arrows indicate induced expression.

The affinity purified GST-fusion protein was tested for its ability to bind to the reported SIX5 binding sequence ARE in a specific manner using gel retardation assays (Fig. 2.7). This ability had previously been demonstrated by Feng Li (personal communication) and was necessary in order to establish the DNA-binding specificity of Six4. If one is to infer the binding specificity of Six4 based on the GST recombinant protein then it is important to demonstrate that the GST fusion protein accurately emulates the Six4 binding specificity.

In the absence of other data regarding the DNA-binding properties of Six4 the ARE enhancer is the only known Six4 binding sequence. A 40 bp long oligo containing the binding sequence found in the ARE enhancer was generated and labelled with 32Pgamma ATP. This oligo (referred to as AREo) was used as a probe in Electrophoretic Mobility Shift Assays (EMSAs) performed with the GST-SD+HD recombinant protein (see section 5.1.2.6 for the AREo sequence, Six5 binding site shown in bold). EMSAs were performed in the presence of 50 ng/µl poly(dI•dC) double stranded carrier to compete with non-specific binding. A competition experiment was performed to confirm the specific nature of ligand binding (Fig. 2.7.2). The retarded band present on the gel is visible in lane 1 but is slowly eclipsed by the progressive addition of 10-, 50- and 100- fold molar excess of unlabelled (cold) AREo probe (lanes 7, 5 and 4 respectively)(see below). No retarded band is observed in the absence of recombinant protein containing WCE (lane 2) or in the presence of WCE from cells transformed with the PGEX-2T plasmid (no insert) and induced under the condition described previously. It was determined that the recombinant protein showed specific affinity to a previously reported Six4 binding site and would therefore be suitable for conducting a SELEX experiment. These findings suggest that binding of the recombinant protein to the sequence of the ARE enhancer occurs in a sequence specific fashion.

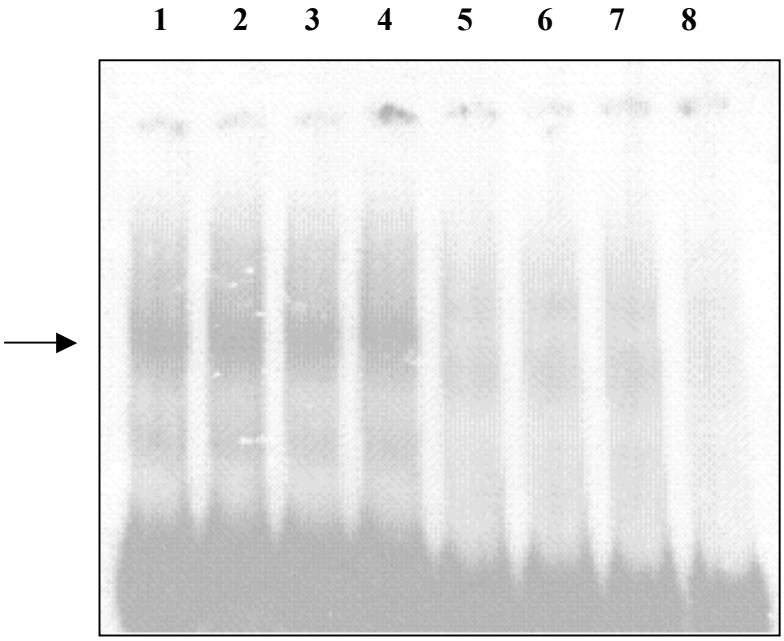


Fig. 2.7.1 EMSA using affinity purified GST-SD+HD recombinant protein**.** Different concentrations of a P^{32} labelled duplex bearing the ARE enhancer Six5 binding element (GGTGTCAGGTTGC) binding sequence (Lanes 1-4, concentrations 10, 5, 2 and 1 fmol respectively) as well as different concentrations of the P^{32} labelled semi-random oligo designated R57 (Lanes 5-8, concentrations 10, 5, 2 and 1 fmol respectively) (section 2.8) were incubated with WCE from IPTG induced *E.coli* cultures expressing the GST-SD+HD recombinant protein (section 2.8). The binding buffer was supplemented with 50 ng/µl poly(dI•dC), 10% glycerol. Black arrow indicates shifted protein-DNA complexes.

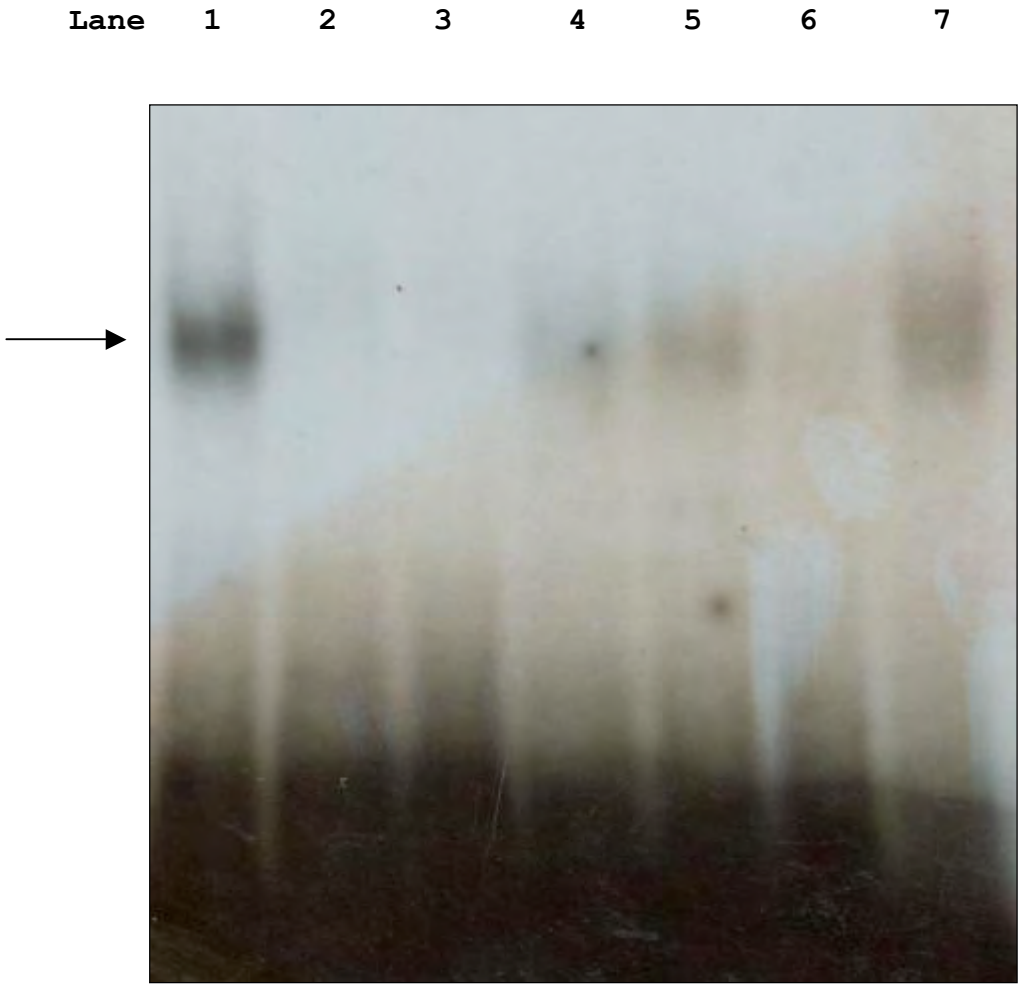


Fig. 2.7.2 Competition EMSA using affinity purified GST-SD+HD recombinant protein**.** 5fmol of a P^{32} labelled AREo are used as the probe in each lane. WCE from IPTG induced *E.coli* cultures expressing the GST-SD+HD recombinant protein (section 2.8) has previously been coincubated with the probe in lanes 1, 4, 5 and 7. Lanes 4, 5 and 7 contain 100-, 50- and 10- fold molar excess of unlabelled (cold) AREo probe. Lanes 2 and 6 contain labelled probe in the absence of WCE. The probe in lane 3 has previously been co-incubated with WCE from cells transformed with the PGEX-2T plasmid (no insert) and induced under the condition described previously. The binding buffer was supplemented with 50 ng/µl poly(dI•dC), 10% glycerol. Black arrow indicates shifted protein-DNA complexes.

2.9 Random Oligonucleotide Pool generation

Originally SELEX was used to select target sequences from genomic DNA libraries. However it quickly became apparent that replacing genomic DNA libraries with semi random oligonucleotides or aptamers (also known as random-mers) greatly enhanced the selective abilities of the SELEX method. Aptamers are smaller, more stable, can be chemically synthesised, and can be radioactively labelled without affecting their affinity for the ligand in question (Rimmele, 2003; Tuerk and Gold, 1990). The aptamers used in this study consist of a random core of a fixed length (no nucleotide bias) flanked by two known ‘‘arms’’ to which primers anneal to allow for PCR amplification.

As discussed previously, SELEX is based on the selection of functional aptamers from a pool of ‘‘random’’ oligonucleotides. As such the initial composition of the pool of oligos, as determined by oligo structure, is a crucial factor in determining the assay’s dynamics and efficiency. According to Marshal and Ellington (2000) (reviewed in Gopinath, 2007) four main factors are involved in oligo pool generation. These are the type of randomization (of the core region), the length of the random sequence region, the chemistry of the pool, and the utility of the constant regions. Of these only the first two are subject to modification since the pool must consist of double stranded DNA and the utility of the constant regions is the PCR amplification of the selected aptamers (their design is discussed below). The core region was completely randomised (no nucleotide bias) since anything but complete randomisation might bias the selection process. Finally, the length of the random oligonucleotide core of the oligos that were used in SELEX was carefully considered during experimental design.

Shorter core regions have the benefit of greater potential representation of each possible permutation of the target sequence given a constant initial reaction stoichiometry. This increases the chances of specific binding and reduces the requirement for numerous selection and amplification rounds. Conversely longer regions, whilst allowing for more possible permutations of the target region, can be expected to contain less representatives of each permutation. They do, however, have the benefit of allowing for the coverage of longer binding sites and the potential inclusion of neighbouring specificity-conferring and/or binding enhancing regions. Given the relative absence of information about Six4 binding it was difficult to reach

an informed decision as to the optimal length of the random core before conducting the experiment.

An additional consideration is the mechanics of the amplification step between rounds. It has previously been reported that the PCR amplification step involved in SELEX may generate undesired by-products (Musheev and Krylov, 2006), potentially, through mispriming of the reaction. This eventuality is much more likely when using a longer random core (for a more complete analysis of this see section 2.11). However, in the absence of any concrete knowledge, the decision of the length of the random core was made based on the methodology utilised by previous studies. Specifically, the variable core sizes vary from 12 as used by Thiesen and Bach (1990) to 15 as used by Xiang et al. (1995), to 19 as used by Lazebnik et al. (2008) to 25 as used by Rami Jarjour to determine the binding specificity of Six5 and by Choi and Sinha (2006) to determine the consensus binding site of ESE-2 (Elf5). At the time the assay was first performed the most commonly used core length was in the range of 25. This size allows for the detection of longer binding sequences and still allows for saturation of the relevant sequence space based on the initial stoichiometry of the SELEX reaction. Figure 2.8 re-summarises the process of SELEX.

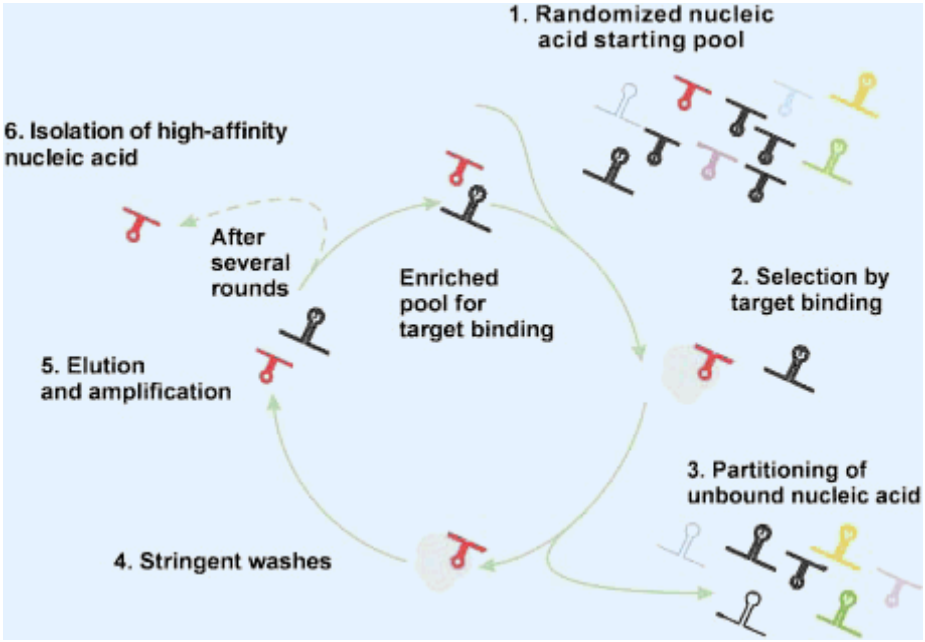


Fig. 2.8 SELEX overview. An initial randomised oligonucleotide pool (1) Target-binding aptamers (2) are segregated from nonbinding oligos (3) because of their affinity to the used ligand (in this case GST-SIX-HD). After washing (4), bound aptamers are eluted from protein they are PCR amplified (5), and used as a starting pool in the next round of selection. After a number of rounds, high-affinity aptamers can be isolated (6). (Figure presented as published in Rimmele, 2003)

It is important that each possible sequence variant to which a utilised protein can bind be represented in a large number in the initial SELEX library. A consequence of the presence of high copy numbers of each sequence variant is that stochastic effects (e.g. loss of sequence variants due to random fluctuations) can be generally neglected in SELEX. These effects typically affect small data sets. When representation of each sequence is small (~10 copies) it is conceivable that binding aptamers may not be recovered during the first round of SELEX due to chance. Once sequences are lost from the selection they can not resurface later and can therefore become extinct. It is therefore imperative that the initial representation of aptamers in SELEX is enough to overcome these effects (Djordjevic and Sengupta, 2006; Levine and Nielsen-Hamilton, 2007; Irvine et al., 1991; Vant-Hull et al., 1998).

Random oligos were initially designed in accordance with previous experiments conducted by Rami Jarjour (personal communication) whilst taking into account the size of previously reported Six5 target sites. The sequences that were reported to show an affinity to the closely related Six5 protein were reported to be 13, 10 and 7 bp long (corresponding to the ARE enhancer sequence, the Six5 SELEX-derived binding sequence and the sequence of the TREX enhancer respectively). The random oligos that were initially used in this study contained a variable core that was 26bp long (compared to 25bp as used by Rami Jarjour). A random core of a size of 26 bp can provide complete coverage of sequence space (of all possible 13 bp long combinations) based on the yield of amplifiable DNA generated by a DNA synthesiser and can potentially include up to 4^{26} possible combinations, although in actuality the number represented in the molecules in the starting SELEX oligo pool is but a fraction of that given the limited concentration of random oligos in the starting reaction (~10^{15} starting molecules). However since the reported Six5 binding sequences are never longer than 13bp the representation of these sequences in the original aptamer pool is much higher since the 13bp recognition sequence can occur in 13 different frames within the random core region. Based on these facts a random oligo core size of 26bp was deemed to be appropriate given both the previously established DNA binding requirements of the murine homologue as well as the stoichiometry of the SELEX reaction.

During the design of these oligos steps were taken to ensure that the flanking primer annealing regions (designated primer 1 and primer 2) of the pool oligonucleotides did not bear a significant similarity to the reported binding sequences and therefore invalidate the results of this assay. Random primer sequences were

generated using the *random sequence* utility at http://rsat.scmbb.ulb.ac.be/rsat/. The nucleotide frequencies utilised in the creation of the random sequences were 0.25 for all four nucleotides. Generated primer sequences were checked for secondary structure, primer dimer formation and melting temperature using the DNA calculator utility at http://www.sigma-genosys.com/calc/DNACalc.asp. The generated primer sequences were subjected to a motif scan using the IUPAC expressions of the reported Six5 target sequences (a hit requires a complete match). Two sequences that returned no hits were selected and used to create the starting oligo library. Their melting temperatures were 54.82° C and 68.9° C for primers 1 and 2 respectively and both sequences tested negative for both secondary structure and primer dimer formation.

The generated pool of oligonucleotides had the sequence

(5'-GTCAGATCTCTTGGCATTN$_{26}$ACTGTCGATGCGGCACTGTC-3′)

and was obtained from Sigma-genosys^{TM}. The base incorporation likelihood for the 26 variable core positions was set to 0.25 for each of the different nucleotides as described previously. This oligonucleotide library was designated R76. Two PCR primer sequences corresponding to the first (top strand) 25 bases and complementary to the last (bottom) 25 bases were chemically synthesized (AmpPrimer1 and AmpPrimer2).

To confirm that the starting library was generated randomly, the R76 oligo library was cloned in the multiple cloning site (MCS) of the TOPO cloning vector. This was done because the flanking arms lie too close to the random core to allow for the priming of a sequencing reaction. Clones were transformed into the DH5α strain of *E.coli* through electroporation. 25 clones of the original library were sequenced using the M13/pUC primer and the average frequency of different nucleotides in each position of the random core was found to be 48 and 52% for C/G and A/T respectively. No identical clones were detected. Based on these findings I concluded that the R76 oligo library was synthesised in accordance to the desired specifications and did not show hints of bias that might invalidate any subsequent SELEX reactions.

2.10 SELEX limitations and considerations

As discussed previously the high complexity of the pools used in SELEX experiments makes it necessary to amplify functional sequences. This amplification step might in itself contain selective processes that could counteract SELEX-induced selection. What follows is an analysis of additional factors that have been reported to

influence the outcome of a SELEX assay but are not related to the ligand-binding properties of aptamers.

It is sometimes the case that individual aptamers do not survive amplification due to extensive secondary structure and may be lost during this step. The PCR step might include selectivity criteria which are difficult or impossible to control by the experimenter. Molecules that have lower ligand-binding affinity than others, might still be ‘‘overrepresented simply because they appear to replicate slightly better during PCR than their competitors’’(Klug and Famulok, 1994).

Klug and Famulok (1994) report that ‘‘the interplay between different kinetic parameters in glutathione binding used in the selection experiment almost certainly results in the bias of the outcome. As currently performed, affinity elution leads to the preferential enrichment of molecules with a high ‘‘dissociation rate’’. This might be one possible explanation why no aptamer which binds significantly tighter than 100 nM has been described to date’’. This was not a particular concern during this study since the binding site of Six4 was not expected to be significantly longer than 10 bp (based on its homology to Six5 and its ability to specifically bind to AREo) and as such was not expected to bind to the recombinant protein tightly enough to prevent its elution.

Buffer conditions can also be crucial for the final result of *in vitro* selection experiments. The same is true for other parameters, such as the elution volume used to remove non-functional molecules, the selection stringency applied, and the pool complexity.

As discussed above, pool size and complete coverage of sequence space is a parameter which has been shown to be critical in SELEX- experiments. It is yet unclear what the optimal size of a randomized sequence or the optimal degree of mutation in a pool of degenerate nucleic acids is (i.e. the size of the variable region). As mentioned above, complete coverage of sequence space can be expected in randomized regions 25 bases long (assuming a binding site that is shorter than 15 bp). However, aptamers selected from pools of this length tend to recruit constant regions from primer binding sites for binding. Very long randomized regions, on the other hand, can result in dimerization or multimerization of individual sequences, which might lead to precipitation of the DNA involved. This phenomenon has been observed by Bartel and Szostak (1993), who solved the problem by immobilizing the pool noncovalently on agarose.

The importance of these considerations became apparent over the course of the initial SELEX experiment where most of the resulting amplified oligonucleotides were shown to contain internal deletions. For a more in-depth analysis of such problems as well as the strategies used to overcome them see sections 2.11-2.14.

2.11 Initial SELEX screen

A random sequence library was generated from the single stranded R76 oligonucleotides by a primer extension reaction using one of the constant arm-annealing amplification primers (AmpPrimer2) in a 20-µl PCR reaction mixture. Double-stranded DNA fragments were gel purified on a 4% agarose gel.

SELEX was initially performed as detailed in section 5.2.6.3 and as summarised here for convenience. Crude cell lysate from BL21 cells expressing the GST-SD+HD protein was co-incubated with glutathione-sepharose beads (Amersham/Pharmacia) to ‘‘load’’ the GST-SD+HD onto the beads to allow for pull-down experiments. Additional proteins present in the cell lysate that had now affinity to glutathione were subsequently removed. Oligo selection was carried out in the presence of the GST-SD+HD loaded sepharose beads. Oligos with affinity to GST-SD+HD were retained and subsequently eluted and collected.

A fraction of the eluate (10 µl) was used for subsequent PCR amplification with 10 ng of each primer per µl and 0.2 mM each deoxynucleoside triphosphate in a 50-µl reaction mixture. PCR was programmed as 50°C for 1 min, 72°C for 1 min, and 94°C for 30 s for 15 cycles. The PCR products were then gel purified. Five rounds of selection were performed, each followed by a PCR amplification step. The final products were digested with *Eco*RI, cloned into the PGEM-2T easy vector, transformed into the DH5α strain of *E.coli*, and sequenced using either the Sp6 or the T7 primers.

Initially, the oligo library was radioactively labelled with 32P in order to establish the percentage of the oligo pool that was selected after each round. Cerenkoff counting of the isolated oligonucleotide DNA before and after SELEX revealed that after the first SELEX round 3.2% of the oligo pool was selected and successfully eluted. This percentage changed to 3.95% and 2% during the third and second rounds respectively and plateaued at ~3.5% for the fourth and fifth rounds hinting towards potential saturation of the utilised protein by binding oligos. This interpretation was reinforced by the fact that an equimolar solution of radioactively labelled AREo

showed a retention rate of 3.2%. The high percentage of oligos selected during the first cycle suggested that the specificity of the utilised protein was low. Given the large error margins associated with these measurements (in the range of 1.2%) it quickly became apparent that monitoring the efficiency of the selection process during the first cycle was uninformative in quantifying DNA selection but did highlight the possibility of increased non-specific ligand binding. I attempted to counter the possibility of non-specific binding by increasing the concentration of poly dI-dC double stranded carrier DNA to 60 µg/ml with little success (2.8% of labelled oligos were selected). An alternative way of avoiding saturation of the ligand would be to decrease the concentration of the starting oligo library. This was avoided, however, since it would also decrease the complexity of the starting library. A reduction of concentration also increased the risk of reducing the yield of selected DNA to such a degree that it would be hard to amplify through PCR. PCR amplification would have to be carried out for more cycles in order to yield the concentration of DNA required to perform subsequent SELEX rounds. This would in turn increase the risk of generating PCR by-products that are common in aptamer amplification (see sections 2.11-2.14). In light of these findings, no alterations were made to the protocol and section was carried out for 5 cycles.

The first SELEX experiment performed yielded a total of 50 potential binding sequences, 27 of which were shown to harbour internal deletions of sizes ranging from 3 to 25 base pairs (for sequences see Fig. 2.9). 23 sequences were found that contained only the invariable flanking arms but no core region and another four (seqs 3.4, 3.5, A.4 and A.5 in Fig. 2.10) harboured only partial deletions. Possible explanations for this occurrence as well as ways to minimise by-product formation are discussed in sections 2.11-2.14. Even though the high turnover of aptamer artefacts was alarming the lack of homogeneity between the remaining full-length aptamers was even more so.

All of the remaining 23 sequences were analysed using the ClustalX and T-Coffee multiple sequence alignment (MSA) utilities (default settings). For the purposes of this analysis only the core regions of the aptamers were considered since the identical flanking arms would disorient any attempt to identify highly similar regions. Additionally, the sequences were also compared manually but no convincing MSA was produced. Indeed the sequences seemed to be very heterogeneous possibly hinting towards a fault in the selective process of SELEX. In addition to looking for an MSA, the 23 sequences were also probed for recurring motifs using the MEME motif elicitation program (default settings, Bailey and Elcan, 1994), all possible motif

sizes between 6 and 13 were used (to account for the sizes of the reported binding sequences of the homologous Six5) but no appreciable similarity was discerned between members of the isolated oligonucleotide pool since no motif was identified that was common to more than 4 out of the 23 (Fig. 2.9). The best identified motif discovered by MEME is shown in Fig. 2.10. The preference for motif elicitation over multiple sequence alignment as well as the considerations involved in both processes are discussed in greater detail in chapter 4.

These findings hinted towards limitations in both the specificity (the ability of the process to recognise genuine Six4 binding sites) and selectivity (the ability to discriminate between specific and non-specific binders) of the SELEX process used but also in the ability of the selected aptamers to be amplified properly. Based on these observations the SELEX approach was revisited in order to address these issues.

```
1.1 CAGGTCAGTTCAGCGGATCCTGTCG ATAGAGTTAAACTAGATGCGGGTTTT GAGGCGAATTCAGTGCAACTGCAGC
1.2 CAGGTCAGTTCAGCGGATCCTGTCG CTTGGNCANGTACTNCATNCANNTGT GAGGCGAATTCAGTGCAACTGCAGC
1.3 CAGGTCAGTTCAGCGGATCCTGTCG ATNANTGACCATAACTAGANGGANTN GAGGCGAATTCAGTGCAACTGCAGC
1.4 CAGGTCAGTTCAGCGGATCCTGTCG TTATGCCATATAATACGATTGGGGTA GAGGCGAATTCAGTGCAACTGCAGC
1.6 CAGGTCAGTTCAGCGGATCCTGTCG GATCCGGTGAATGACTGAACATTGAC GAGGCGAATTCAGTGCAACTGCAGC
1.7 CAGGTCAGTTCAGCGGATCCTGTCG AACTGCTTCCATGGAAGTGTGACAAC GAGGCGAATTCAGTGCAACTGCAGC
1.8 GAGGTCAGTTCAGCGGATCCTGTCG  AAGAAACAAGGTAGGAAGGCAGATA GAGGCGAATTCAGTGCAACTGCAGC
2.3 CAGGTCAGTTCAGCGGATCCTGTCG TACTCCNAATATAATGTCCCCGCNAG GAGGCGAATTCAGTGCAACTGCAGC
2.4 CAGGTCAGTTCAGCGGATCCTGTCG TTACCGCCTGNACCTGTANGNTACTT GAGGCGAATTCAGTGCAACTGCAGC
2.5 CAGGTCAGTTTCAGCGGGNTCCCGN TGGTTCACCACCATCGTGTANNTCAG ANAGGAGGCGAATTCAGTGCAACTGCAGC
2.6 CAGGTCAGTTCAGCGGATCCTGTCG GGAAGTCAAGATCGTTGAAGCTAAGA GAGGCGAATTCAGTGCAACTGCAGC
2.8 CAGGTCAGTTCAGCGGATCCTGTCG ACAATCCACTCGTAGGGATTGTACTT GAGGCGAATTCAGTGCAACTGCAGC
2,9 CAGGTCAGTTCAGCGGATCCTGTCG ATGTCTATCATGCATCAATGCGTGAT GAGGCGAATTCAGTGCAACTGCAGC
2.10CAGGTCAGTTCAGCGGATCCTGTCG CGTAACGATATCAGGTAGCGATCTCA GAGGCGAATTCAGTGCAACTGCAGC
3.1 CAGGTCAGTTACAGCGGATCNTGTT GNATCTTTAGTATTCTGNAAAATNAG GAGGCGAATTCAGTGCAACTGCAGC
3.2_CAGGTCAGTTCAGCGGATCCTGTCG AGAGCGATTTATTCATCTGGAGCTTT GAGGCGAATTCAGTGCAACTGCAGC
3.3 CAGGTCAGTTCAGCGGATCCTGTCG TCCCCCGTGACTCCAACTAGCATATT GAGGCGAATTCAGTGCAACTGCAGC
3.4 CAGGTCAGTTCAGCGGATCCTGTCG                       GCAAA GAGGCGAATTCAGTGCAACTGCAGC
3.5 CAGGTC                  TGGG CAATGTT              GAGGCGAATTCAGTGCAACTGCAGC
3.6 CAGGTCAGTTCAGCGGATCCTGTCG TCAGGTCAGTTCACCCGGATCCTGTC GAGGCGAATTCAGTGCAACTGCAGC
3.8 CAGGTCAGTTCAGCGGATCCTGTCG ACACAATGTAGTTCAGCTGTGGGTCA GAGGCGAATTCAGTGCAACTGCAGC`
3.10CAGGTCAGTTCAGCGGATCCTGTCG AAACGTAACACTAGACATGACTCGGA GAGGCGAATTCAGTGCAACTGCAGC
4.2 CAGGTCAGTTCAGCGGATCCTGTCG GATAGTGAGATATGGTAATGTATGTA GAGGCGAATTCAGTGCAACTGCAGC
4.6 CAGGTCAGTTCAGCGGATCCTGTCG AAGGTAGGTCAATCCACACTCGCACGCGAGGCGAATTCAGTGCAACTGCAGC
4.7 CAGGTCAGTTCAGCGGATCCTGTCG GAAAAGAATTGATCAGGGCTGGCAGGAGAGGCGAATTCAGTGCAACTGCAGC
A.4 CAGGTCAGTTCAGCGGATCCTGTCG                     ACAGG AGNCGAATTCAGCTGCAACTGCAGC
A.5 CAGGTCAGTTCAGCGGATCCTGTNG                   GATATA GAGGCGAATTCAGTGCAACTGCAGC
```

Fig. 2.9 Multiple alignment of the sequences of the isolates obtained from the initial SELEX screen. Gaps indicate missing sections of the original R76 oligonucleotide pool the isolates originated from. These truncations were attributed to PCR mis-priming and/or improper amplification. Sequences encountered in multiple copies are only represented once in this alignment. For convenience all sequences from which the internal variable region was completely deleted were omitted. The core sequences are underlined.

```
AGGTCAGTTC ACCCGGATCCTGT C
     GGTTC ACCACCATCGTGT ANNTCAGA
       AAA ACCCGCATCTAGT TTAACTCTAT
         T ACCCCAATCGTAT TATATGGCAT
```

Fig. 2.10 Multiple alignment of 4 of the 23 sequences isolated through the initial SELEX experiment. The motif identified in all 4 sequences by the MEME motif elicitation utility (Bailey and Elcan., 1994) is shown in bold.

2.12 Generation of undesired truncations during PCR amplification

It was decided that the truncation generating forces and the lack of specificity and selectivity were both issues that needed to be addressed lest they invalidate any conclusions drawn by this study. The problem of SELEX by-product generation is not unheard of. Musheev and Krylov (2006) published a study of the PCR amplification of random DNA libraries used in aptamer selection. They report that capillary electrophoresis had identified ‘‘fundamental differences between PCR amplification of homogeneous DNA templates and that of large libraries of random DNA sequences’’ (Musheev and Krylov, 2006). Specifically, when a random DNA library is used as a template for PCR, product accumulation stops when PCR primers are still in excess of the products. The products then rapidly convert to what was characterised as by-products (not to be confused with non-specific PCR products) and virtually disappear after only 5 additional cycles of PCR. The yield of the products was shown to be inversely correlated to the increasing length of the DNA molecules in the library. It was additionally reported that the initial number of DNA molecules in a PCR mixture has no effect on the rate of by-product formation. Similarly, the increase of the *Taq* DNA polymerase concentration in PCR mixtures selectively increased the yield of PCR products. They concluded that ‘‘standard procedures of PCR amplification of homogeneous DNA samples cannot be transferred to PCR amplification of random DNA libraries: to ensure efficient SELEX, PCR has to be optimized for the amplification of random DNA libraries’’ (Musheev and Krylov, 2006).

According to Musheev and Krylov (2006), by-products are ss-dsDNA, which are formed through product–product hybridization. By-product formation therefore only starts when a threshold amount of the products is generated. This was empirically determined to be 20–50 nM (total of 10^{11}–10^{12} molecules) in a reaction of comparable stoichiometry to the one described in this study and for a pool composed of 88bp long oligos (as opposed to 76 in the study presented here).

These reports highlighted the fact that since by-products were being produced during PCR then the pool of template DNA must have a high degree of heterogeneity. This in turn points towards limited selectivity in the SELEX reaction. If binding during SELEX occurs in a non-specific fashion then the heterogeneity of the selected aptamers will drive by-product formation.

Musheev and Krylov (2006), however, are rather vague about the specific events that drive by-product generation. I suspect that these truncated products occur because of preferential amplification of randomly occurring oligos during the PCR reaction. It is conceivable that a sequence bearing a sufficient similarity to the primer arms could randomly occur within the variable oligonucleotide core. If the affinity of this sequence to the primers is sufficient to cause the amplification of one such oligo then the incorporation of the primer into the next product will create a product of reduced length that contains exact matches to both primers. If such a product is then preferentially amplified during the PCR reaction due to its reduced length then its concentration will increase exponentially.

If this theory holds true then an exact match to at least the last 18bp of the known primers would be required for the oligo to achieve the minimum base-stacking calculated T_m of 50^0 and thus be amplified. This fact is based on an analysis that was carried out using an electronic T_m calculator (http://www.promega.com/biomath/). Additionally, the replication time of the truncated PCR artefact would have to be significantly smaller than that of the full length oligos if preferential amplification is to occur.

Deletions represent a far larger portion of the sample than these assumptions would explain. If this theory is to be believed then truncated oligos should be significantly favoured in the PCR amplification reaction. Their inclusion level of 56% (27/50) however can not be explained based on the estimated starting reaction frequency of oligos that match to either of the primer sequences on 18 or more positions. Additionally the complete exclusion of the random core region of the oligo would only reduce the replication time by 1.5secs or 2.5% of the total extension time allotted in one PCR cycle (assuming synthesis of one kb of DNA occurs over 1 minute).

Another possible explanation for this phenomenon would be the presence of hairpin loops within the random core of the oligos that would preclude part of the core from being replicated by polymerase and thus resulting in a truncation event. Theoretically the temperatures involved in the PCR reaction would linearise these structures although such events have been known to occasionally occur (Andrew Jarman and Ian Simpson, personal communication).

It was therefore, initially, thought that a certain lack of selectivity can be expected in the first round of SELEX given the fact that ‘‘strong’’ binders do not exist in large enough numbers in the original aptamer library so as to dominate the population of

selected aptamers during the first cycle. This in turn could mean that weaker, non-specific binders may associate with the ligand and provide the heterogeneity required for by-product formation. It was therefore deemed imperative that by-product formation be stopped through optimisation of the PCR amplification so as to allow sufficient enrichment of strong binders in subsequent SELEX rounds so they could later over-take the population.

This problem was initially addressed through optimisation of the PCR process through altering of the buffer conditions through the use of the Opti-Prime™ PCR optimisation kit (Stratagene). The Opti-Prime™ PCR optimisation kit allows for the assessment of the fidelity and the efficiency of a PCR reaction through the use of a number of different buffering conditions. 12 amplification reactions of the R76 oligo library as well of a single oligonucleotide selected in the original PCR (oligo 1.1) were performed in parallel. The reaction conditions used were those suggested by the Optimal buffer determination protocol and essentially do not differ from those described in section 5.2.6.2. The 12 different reaction sets utilized the 12 different PCR buffers supplied with the Opti-prime kit (the final reaction concentrations of the various buffer components can be found in appendix 2.1). PCR fidelity optimization considerations are discussed below (Section 2.13) for convenience. The amplified DNA was then run on a 4% low melting point agarose gel to assess the length of the amplified oligos. Additionally, the amplified oligos were cloned into *E.coli* and 3 clones from each reaction in the case of R76 and 1 clone in the case of 1.1 were sequenced as described in section 2.11. No truncations comparable to those that occurred during the original SELEX experiment were observed when oligo 1.1 was amplified and all sequenced clones were found to be 76bp long. Conversely, 23 out of 36 sequencing reactions involving the R76 oligo pool showed truncations ranging from 5 to 20 bp. All gel bands originating from the R76 oligo library appeared indistinct and roughly co-migrated with the 50bp fragment on the marker ladder in contrast to those originating from 1.1 that appeared as a single band that co-migrates with the original R76 oligo library.

These findings are consistent with those of Musheev and Krylov (2006) as well as with the theories proposed previously. Specifically it was apparent that aptamer pools of high variability (R76) were prone to being overtaken by truncated by-products during PCR amplification. This was not true of PCRs using homogeneous templates (1.1). This also highlights the inability of the initial SELEX cycle to reduce the oligo library from a heterogeneous selection of aptamers to a much more homogeneous

mixture through ligand-binding mediated selection. These issues are addressed in sections 2.14 onwards.

2.13 PCR Fidelity optimisation

The fidelity of PCR is dependent on several variables. These include, the concentrations of dNTPs and magnesium in the amplification reaction. Innis et al. (1988) report that lower initial concentrations of dNTPs (between 20 and 100 μM), as well as lower magnesium concentrations can result in ‘‘greater fidelity and specificity due to lower misincorporation of nucleotides’’. The final dNTP concentration recommended with the Opti-Prime PCR optimization kit is 200μM, but lower concentrations were also used.

Finally, the number of cycles will affect the fidelity of the final product, because PCR products resulting from misincorporated nucleotides will serve as templates for further extension and proceed exponentially from the point of the first *false* PCR product. Since high fidelity was a concern in the cloning of PCR products the number of temperature cycles was limited to the minimum required to give the necessary amount of DNA for the procedure (15 cycles were used in this case). Finally the presence of DMSO in the reaction mix has been known to reduce the stringency of base pairing during replication and therefore potentially allow for annealing of primers within the central variable region.

Despite using a wide range of parameters when carrying out the amplification step the tendency for the core aptamer regions to be truncated persisted. This problem in turn prompted a revision of the experimental approach.

2.14 Revised recombinant protein design

In light of the findings of the initial SELEX experiment a wide range of parameters involved in the experimental procedure were systematically revisited. Chief amongst those was the design of the recombinant GST-SD+HD protein since the generation of by-products suggests a potential lack of selectivity on behalf of the protein utilised. A more in-depth analysis of the SIX and homeodomains of Six4 reveals a discrepancy in the topology of the homeodomain as it was previously reported by Seo et al. 1999 (positions 296-351, Fig. 2.11) A secondary structure analysis of the Six4 homeo- and SIX domains conducted using the Prof secondary structure prediction utility (http://www.aber.ac.uk/~phiwww/prof/, Ouali and King,

2000) revealed that the predicted termination point end of the homeodomain as defined by Seo et al. (1999) occurred within a predicted coil structure (confidence value of 0.85 and 0.95 for all 6 coiled residues positions 349-355). These predictions were performed based on the homology between the Six4 homeodomain and other known homeodomains and essentially utilise the same methodology used by Seo et al. (1999). The ramifications of this were that the resulting recombinant protein that was designed and expressed using this topology may be incomplete and therefore have an altered or compromised specificity when compared to the full length Six4 protein therefore accounting for the apparent lack of specificity displayed by the protein utilised in the initial SELEX screen. According to the predicted structure the end of terminus of the homeodomain may lie as far as position 355 compared to a previously reported 351, suggesting a potential exclusion of at least 4 amino-acids from the original recombinant protein. Additionally, the protein domain recognition program PROSITE (http://www.expasy.ch/prosite/, Hulo et al., 2006) reports that the topology of the homeodomain of Six4 further differs from that reported by Seo et al. (1999) and determines its topology as 303 – 354.

It is conceivable that the exclusion of these amino acids may cause an alteration in the binding specificity of the resulting protein. Similar results have been observed when mutating certain nucleotides within the homeodomain encoding region of the *HSF1* gene from *Saccharomyces cerevisiae* that severely compromise the ability of HSF to bind to its normal binding site, repeats of the module nGAAn. One of these mutations, Q229R, shows a ‘‘new specificity’’ phenotype, in which the protein prefers the mutant sequence nGACn. These results identify the region of HSF that contacts DNA (Torres and Bonner, 1995). These findings may have similar implications for the DNA-contacting and specificity-conferring residues of the Six4 homeodomain. Even if the topology defined by Seo et al. (1999) does not compromise the specificity of the resulting recombinant protein, its use would seriously weaken any conclusions reached by this study.

This realisation necessitated the redesign of the GST-SD+HD recombinant protein. The sequence corresponding to the SIX and homeodomains of Six4 as defined in the revised topology was amplified from cDNA using new primers (for sequences see section 5.1.2.6). The resulting fragment was then expressed as a recombinant GST-fusion protein in the pGEX-2T expression system. The redesigned protein incorporated regions of the homeodomain extending to position 365 and was expected to be biochemically functional based on secondary structure predictions

(Fig. 2.11). The affinity purified GST-fusion protein was tested for its ability to bind to the reported SIX5 binding sequence AREo using gel retardation assays (Fig. 2.7.a). Shifted bands were observed in the presence of radiolabelled AREo. The bands were progressively phased out by the addition of 10-, 100- and 200-fold molar excess of unlabelled AREo. Conversely no retarded bands were observed in the presence of radiolabelled R76. These findings suggest that the new recombinant protein binds to AREo in a specific manner and is therefore, potentially representative of the DNA-binding specificity of Six4. This protein was then used in a new SELEX experiment.

MFDKNLDGNNLSVSIGGDLDSTSSGG
CCCTTCTTCCEEEEEECCCCCCCSSC
TSSDHSAVHQDNLSSPMAYGSLFLPNAGYRGNLSCKTVLQLDKFAPYEGVEKDHLLERRFQDITND
CCCCCEHHEHTTCCCHHCHEEEECCTTTCCSCCCHHHHEEECTTCCCTTCHHHHHHHHHHHHHHC
YDKSPPPTASTTPTHYPSLNSIIFENGSSGNLGDLNGNTKTDLCAGLQRSGGGLGGNAGSGGHLIS
TCSCCCCCCCCCCCCCCCCCEEEEETSCCSCCTCCTTCCHHHHHHHHEETTCCCCCCCTCHHHHHH
NLTAAHNMSAVSSFPIDAKMLQFSTDQIQCMCEALQQKGDIEKLTTFLCSLPPSEFFKTNESVLRA
HHHHHHHCEEEECCHHCHHHHTCCHHHHHHHHHHHHHHTTCHHHHHHHHHHCCCHHHHHHHHHHHH
RAMVAYNLGQFHELYNLLETHCFSIKYHVDLQNLWFKAHYKEAEKVRGRPLGAVDKYRLRKKYPLP
HHHHHHHHHHHHHHHHHHHHCCCCHTHHHHHHHHHHHHHHHHHHHHTTCCCCCCCEEEEEECCCCC
KTIWDGEETVYCFKEKSRNALKDCYLTNRYPTPDEKKTLAKKTGLTLTQVSNWFKNRRQRDRTPQQ
CEEECTCEEEEEEHHHHHHHHHHHHHHCCCCCHHHHHHHHHHTTCCEEEHHHHHHHHHHHHCTCC
RPDIMSVLPVGQLDGNGFPRMFNAPSYYPETIFNGQ
CCCEEEECCEECETTTSCCEEECCCCCCCEEEETCC

Fig. 2.11 Prediction of the secondary structure of the full length Six4 protein. Six4 Primary sequence structure shown in Italics and in bold above the predicted secondary sequence structure (8 class). Figure key: H=alpha helix, E=beta strand, C=coil, T=turn. Underlined regions represent the SIX-homeodomain that were incorporated in the GST-SD+HD recombinant protein. Region in red is excluded from the topology determined by Seo et al. (1999) but is included in the revised recombinant protein designated GST-SD+HD+.

2.15 Revised SELEX screen findings

SELEX was re-performed using the same parameters utilised in the initial screen (See sections 2.11 and 2.14), with the only difference being the use of the redesigned GST-SD+HD protein (designated GST-SD+HD+).

After 5 SELEX rounds 65 isolates were amplified cloned and sequenced as described previously. 59 (92%) of these were found to harbour almost complete deletions of the central variable region. The sequences of the remaining 6 aptamers can be seen in Fig. 2.12. No discernible similarity was observed between these sequences (for sequence analysis methodology see section 2.11). Additionally, none

of the isolates showed any similarity to previously reported SIX4/5 subfamily binding sequences although 1 (designated B.2) was found (through MEME analysis using the parameters described in section 2.11) to be similar to isolates recovered in the initial SELEX experiment (Fig. 2.13). This apparent heterogeneity in the selected aptamers could potentially be attributed to the disruptive effects of by-product formation. Therefore no further analysis of the selected oligos took place until the by-product problem was addressed.

The prevalence of artefacts in the pool of isolates was confirmed through agarose gel electrophoresis of the entire pool (Fig. 2.14). Once again the electrophoretic mobility of the pools of isolates resulting from the last three rounds of SELEX was found to be higher than that of the oligonucleotide pools they originated from suggesting the oligos contained therein were shorter than expected. The recurrence of this phenomenon suggested that the problems inherent in the amplification process could not be addressed by the redesign of the recombinant protein alone and that other steps needed to be taken to ensure the success of SELEX. This phenomenon was attributed to an inherent fault in the design of the aptamer library. In light of this the aptamer library was redesigned to limit the effect of truncation inducing cues as speculated in section 2.11.

```
>5/3 CAGGTCAGTTCAGCGGATCCTGTCG AAATTCTATACATTTCGATTTAATCT  GAGGCGAATTCAGTGCAACTGCAGC
>4   CAGGTCAGTTCAGCGGATCCTGTCG ATTCCNCAGTGATTTCNNCCCGCTTGA GAGGCGAATTCAGTGCAACTGCAGC
>1   CAGGTCAGTTCAGCGGATCCTGTCG TTTTGCTATTCTTACAATTGGTATATG GAGGCGAATTCAGTGCAACTGCAGC
>15  CAGGTCAGTTCAGCGGATCCTGTCG ATACAAAATGTAATTTGACACATTTTG GAGGCGAATTCAGTGCAACTGCAGC
>10  CAGGTCAGTTCAGCGGATCCTGTCG CAGTTGCACTGAATTCGCCTCTACTCNTATCAAATANCCNGGAGGCGAATTCA
GTGCAACTGCAGC
>9   CAGGTCAGTTCAGCGGATCCTGTCG TAACGTAATCATAATCTAAGCTAGTTG GAGGCGAATTCAGTGCAACTGCAGC
```

Fig. 2.12 Multiple alignment of the sequences of the isolates obtained from the second SELEX screen. Gaps indicate missing sections of the original R76 oligonucleotide pool the isolates originated from. These truncations were attributed to PCR mis-priming and/or improper amplification. Truncated isolates are omitted from this alignment.

```
2.10        TGAG ATCGCTACCTG ATATCGTTAC
B.2   CGCCTAAGAT ATCGCTAAGTG TAC
2.8     AAGTACA ATCCCTACGAG TGGATTGT
```

Fig. 2.13 Multiple alignment and sequence comparison between the variable core region of the isolate resulting from the revised SELEX assay and the GST-SD+HD binding sequences identified in the initial SELEX assay. Sequences 2.8 and 2.10 were isolated in the original SELEX experiment whereas sequence B.2 originates from the revised SELEX.

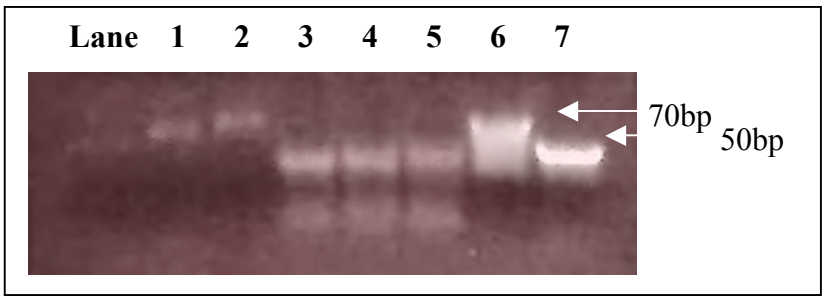


Fig. 2.14 Agarose gel electrophoresis of PCR amplified selected oligonucleotides isolated through SELEX. Lane 2: Double stranded R76, amplified SELEX oligos after 1,2,3 and 4 rounds of selection (lanes 1,5,4 and 3 respectively) exhibit a tendency to decrease in size with each SELEX cycle. Lanes 6 and 7 contain oligos of sizes corresponding to 70 and 50bp respectively.

2.16 Revised oligonucleotide pool design

The need to minimise the truncation generating forces inherent in the PCR amplification necessitated the understanding of those selfsame forces as well as the redesign of the oligonucleotide pool they afflicted. Possible reasons for the generations of lesions have been discussed in section 2.12 but can be summarised as either PCR mispriming or secondary structure generation within the central variable region of the oligo. A possible solution to this problem would be the generation of an oligo with a shorter variable region which would therefore severely reduce the potential for mispriming or hairpin formation through the presence of less potential association-generating bases.

As a way of investigating this theory 3 new variable oligonucleotide pools were generated with the following sequence:

5'-gttagacagtgccgcatcgacagt N_x cgaacgcaatgccaagagatctgac-3' ,

where x took the values of 7,10 and 13 corresponding to the sequence lengths of the previously reported SIX5 binding sequences TREX, minimal ARE and complete ARE respectively (for a discussion on these see section 2.2). This decision was made based on the assumption that given the sequence similarities displayed between the amino acid sequences of the murine and *Drosophila* SIX proteins, a severe diversification of their respective binding sequences to the extent that they would differ in length would be deemed highly unlikely. These oligos were designated R57, R60 and R63 based on their respective sizes. Flanking arms were designed *de novo* using the process outlined in section 2.9.

The reduced length of the core however was expected to directly increase the concentration of every possible permutation of a given length target sequence (for additional information on how this was derived see section 2.9) given a constant reaction stoichiometry. This in turn would theoretically increase the concentrations of sequences with an affinity to the GST-SD+HD+ recombinant protein and as such the concentration of sequences that would present in the beginning of the post-SELEX amplification and therefore increase the homogeneity of the selected aptamers. The reduced length did however mean that potentially longer binding sequences would not be covered by the variable core. This trade-off was deemed acceptable to eliminate the lesion-inducing forces exerted in PCR.

Finally, the inclusion rate of what would be later deemed to be true binding sequences within the pools of isolates obtained through SELEX would also act as an indication of the selective pressure the GST-SD+HD+ recombinant protein would be able to exert on the oligonucleotide pool given the fact that equal initial concentrations of the three different oligo pools would contain different proportions of the expected 7bp long essential binding site core. Therefore an equal representation of the subsequently established Six4 binding sequence in pools of isolates derived from different oligo pools and therefore from different initial concentrations of that selfsame binding sequence would indicate a saturation of the oligonucleotide selection pool at some point in time prior to the conclusion of the SELEX screening process. This could provide a measure of the number of SELEX rounds that need to be conducted in order to accurately establish a binding sequence.

2.17 Final SELEX

The SELEX screen was re-performed in triplicate using the same parameters utilised in the initial screen (See sections 2.11 and 2.15), using the GST-SD+HD+ recombinant protein and oligonucleotides R57, R60 and R63 as the starting aptamer libraries. Agarose gel electrophoresis of the amplified aptamer pools isolated after each round revealed no changes in the mobility of the amplified DNA when compared to the original aptamer libraries, indicating that no truncations were induced. After 5 cycles of SELEX selected aptamers were cloned, transformed, extracted and sequenced as described in section 2.11.

DNA sequences were obtained for a total of 41 isolates from the final round of SELEX, 18 for the R57 aptamer library, 21 isolates from the R60 library and 2 isolates from the R63 library (the R63 library clones proved particularly difficult to sequence). Sequences can be found on table 2.3. Aptamers from all three libraries showed much higher homogeneity than those isolated in previous SELEX experiments with similarities between sequences being immediately apparent.

The information contained within this sequence library will be briefly presented here. Chapter 3 deals with the derivation of a Six4 consensus binding site from the SELEX data presented here in much greater detail and discusses the computational considerations of inferring the binding specificity of Six4 from a multiple sequence alignment (MSA) of SELEX aptamers. The merits of different algorithms and the reasoning behind the utilised methodology are also presented therein.

All the aptamer sequences isolated from the final SELEX screen were probed for recurring motifs using the MEME motif elicitation utility. A 9bp long motif common to 17 (41%) of the isolated sequences was used as a basis for constructing a MSA that was used to inform a position specific scoring matrix (PSSM). The resulting PSSM was used to scan the putative regulatory regions of all identified *Drosophila melanogaster* genes. A graphical representation of this PSSM (expressed as a WEBLogo image) can be seen in Fig. 2.15. Additionally Fig. 2.15 outlines the possible expressions of all Six4 binding sequence positions as well as shows its expression in the IUPAC code (GBHACMBGW). The inferred Six4 binding site consensus sequence contained an exact match to the previously reported TREX binding sequence (Fig. 2.1) although some of the variable positions were found to be more restricted.

The resulting consensus binding site (GTAACCTGA) is highly similar to those previously reported for Six5 (see table 2.5). This fact supports the veracity of the findings of this study since the high degree of homology between homeodomains (as discussed in chapter 1) as well as the affinity of Six4 for the ARE enhancer sequence hint towards a similarity between the binding sequences of these two proteins. The implications of this discovery on the potential conservation of the role of SIX4/5 family members between *Drosophila* and higher eukaryotes are discussed in the following section.

The resulting PSSM was also used as a starting point for conducting a consensus mutagenesis in order to establish the position occupancy requirements of the Six4 binding site. The following sections describe the mutagenesis of the Six4bs and assume the knowledge of its consensus sequence. The computational aspect of this study is dealt with, separately, in chapter 3.

R57 library sequences	R60 library sequences	R63 library sequences
57/2 ATATACT	60/3 GAGTATATCG	63/12 GGTCACCGGACAC
57/17 GATTACA	60/2 CACCTGACAC	63/32 TGAATGCGTTGGA
57/7 TTATATC	60/1x TAACCTGACA	
57/13 ATTCATA	60/5x GATGCCGAACG	
57/11 ATAACATAC	60/10 TATTCGACAC	
57/19 TGTAACCCGA	60/14 CTCGGGTTAC	
57/33 GAAAGCG	60/13 ATTATGTAAC	
57/32 GTTACTT	60/21 GCAACCCGAT	
57/31 AATATCA	60/20 GTCGGGTTGC	
57/30 CAAACCT	60/11 CACCTGACAC	
57/29 ATATTAT	60/19 GTGTCGGGTA	
57/28 GTACGCC	60/18 GATCAGGTTA	
57/27 GCGTCAA	60/9 GGTACATGAT	
57/x GTGGCGG	60/5 TTCGGGTTAC	
57/20 GTCGAAC	60/17 CTCATGTTAC	
57/22 CATCATC	60/8 GTAGACGTGT	
57/23 CGCCGTA	60/15 GTAACCTGA	
57/25 ATATATC	60/16 TCGATGCGGC	
	60/4 AACCGGAAAC	
	60/3x CGATATACTT	
	60/1 GTCAGGTTAC	

Table 2.3 Table of sequences of all the recovered aptamer cores. Sequences are segregated based on the aptamer library they originate from (R57, R60 and R63).

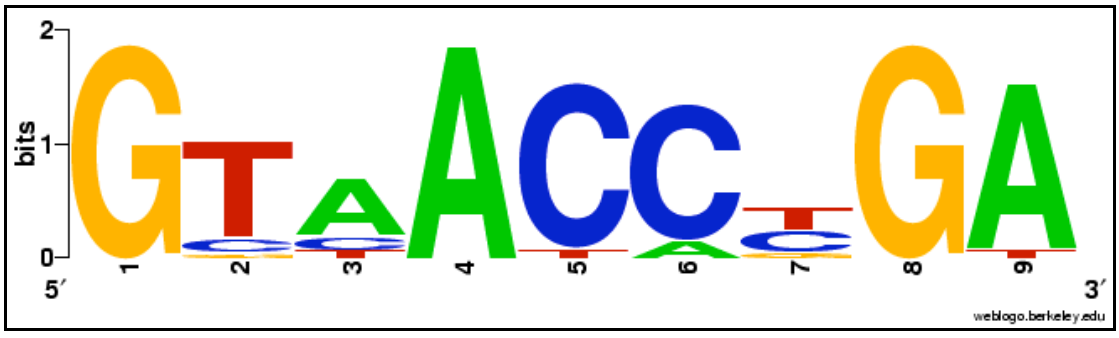


Fig. 2.15 Graphical representation of Six4 consensus binding sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each nucleotide at that position. Observed frequency is expressed here as information content (measured in bits, see section 3.6.2) Image generated using WEBLOGO (weblogo.berkeley.edu, Crooks et al., 2004)

2.18 Verification of specific binding to the consensus binding site (GTAACCTGA)

The specificity of the binding of recombinant GST-SD+HD protein to the Six4bss identified in the SELEX screen, was tested by gel shift assay (EMSA). Affinity purified GST-SD+HD+ protein isolated from crude cell lysate from bacteria expressing the GST-SD+HD and P^{32} labelled oligonucleotides containing the Six4bss flanked by the R57 primer annealing arms (oligo designated as 60/12) were mixed with 30 fold molar excess of unlabelled poly dI-dC double stranded carrier as competitor (Fig. 2.16).

The quantity of the non specific competitor DNA is highly important. If the probe is the only DNA present, any DNA binding protein will tend to bind it. However, if cold DNA that doesn't contain the binding site in question is added, the selective protein will remain bound to the probe, while any other DNA binding proteins present will be spread out over the entire DNA content of the reaction, reducing the non specific retardation of the probe. This fact was of particular importance since previous attempts to purify GST-SD+HD+ that was previously bound to glutathione sepharose beads through glutathione elution showed vastly reduced protein yields and soluble whole cell extracts (WCE) had to be used in EMSAs.

The Six4bss-containing labelled oligonucleotide showed an affinity to GST-SD+HD+ even in the presence of 30 fold molar excess of poly dI-dC. Conversely, the lane containing GST-SD+HD+ and P^{32} labelled R60 showed no shifted complexes thus confirming the fact that ligand binding to 60/12 is specific. This test was an initial assessment of the binding dynamics of GST-SD+HD+ to the Six4bss consensus and was to provide the base of a positional mutagenesis of Six4bss. A more exhaustive analysis can be seen below. The reaction conditions utilised in the EMSA described here as well as in the following mutagenesis analysis can be found in Chapter 5 (5.2.2.16).

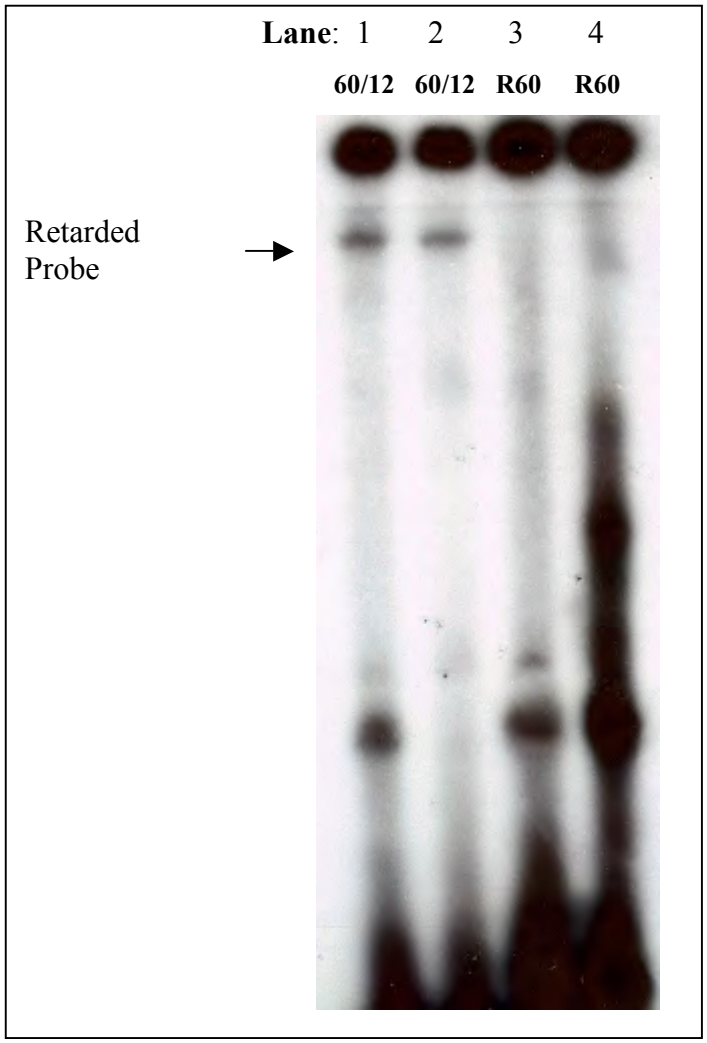


Fig. 2.16 Electrophoretic mobility shift assay of the P^{32} labelled Six4bss-containing R57. **Lane 1**: WCE of BL21 cells containing GST-SIX+HD+ plasmid induced with 1mM IPTG. **Lane2**: WCE of BL21 cells containing GST-SIX+HD+ plasmid induced with 1mM IPTG stored at 4^0 for 2 days. **Lane 3**: WCE of uninduced BL21 cells containing GST-SIX+HD+ plasmid. **Lane 4**: P^{32} labelled R57 incubated in the absence of WCE. Black arrow indicates shifted protein-DNA complexes.

2.19 Binding Sequence Position Occupancy analysis

In order to rule out the possibility of bias in the base occupancy frequencies of each position of the Six4bss due to the potential exclusion of binding sequences from the pool of isolates through chance, Six4bss was mutagenised and tested for binding to GST-SD+HD+. As described previously, the SELEX process is robust against stochastic problems that may affect the initial aptamer library. SELEX, however, is a procedure affected by numerous parameters and an independent corroboration of the results is often useful. This necessitated the assessment of the effects of mutagenising key positions Six4bss on ligand-binding.

All the positions of Six4bss were mutagenised individually so as to include all the nucleotides not present in the SELEX derived consensus sequence (with the exception of the first 2). In the case of the 4 invariable positions the nucleotide transitions were kept within the same nucleotide sub-class (i.e. purine to purine). To keep the altered variables to a minimum, oligonuleotides including the same flanking arms used in the 3rd SELEX experiment as well as the desired consensus permutation were generated in a DNA synthesiser (WMG oligo synthesis) (for complete sequences see table 2.4) to ensure minimal alteration of the binding reaction conditions.

The binding oligonucleotides were rendered double-stranded by a primer extension reaction carried out with the purified 57-mers as template and the bottom-strand primer in a 20-µl PCR reaction, carried out for a single round of amplification.

The oligonucleotides used in the positional occupancy analysis all have the:

cgttagacagtgccgcatcgacag N_7 cgaacgcaatgccaagagatctgac

sequence where the variable central region is replaced with different mutations of the Six4 consensus binding sequence as shown in table 2.4. There were 13 variants that were collectively designated Six4bssmut.

The different Six4bssmut variants were then subjected to relative *in vitro* binding analysis in order to establish their affinity to GST-SD+HD+. Relative affinities of the Six4bssmut variants were established through EMSA analysis of binding of radiolabelled oligos to GST-SD+HD+ (Fig. 2.17). The results of the analysis are presented in Fig. 2.17.

5 of the 13 Six4bssmut variants were found to still have an affinity to GST-SD+HD+, whereas 8 single nucleotide substitution were found to be sufficient to

abolish GST-SD+HD+ binding. Notably Position 1 was shown to be completely variable although a strong tendency towards occupation of that position by Adenine was displayed during SELEX. Position 2, which was previously thought to be invariable, was mutated to another purine without compromising GST-SD+HD+ binding affinity. Likewise position 3 could be mutated to another pyrimidine without affecting binding. Position 4 could support any base apart from Adenine and position 5 was found to be able to accommodate all 4 bases. These results supplement those obtained through SELEX but do not contribute towards informing the Six4 PSSM since the data resulting from the mutagenesis is qualitative rather than quantitative and gives no indication of the effect a single base mutagenesis has on the relative ligand binding affinity of a binding site. This analysis is by no means exhaustive. Quantification of the relative binding affinities of different Six4bss variants may provide additional information about Six4 binding specificity. These considerations as well as others are discussed in greater detail in chapter 3.

core1	cgttagacagtgccgcatcgacagt **g**acctga cgaacgcaatgccaagagatctgac
core2	cgttagacagtgccgcatcgacagt a**g**cctga cgaacgcaatgccaagagatctgac
core3	cgttagacagtgccgcatcgacagt aa**t**ctga cgaacgcaatgccaagagatctgac
core4	cgttagacagtgccgcatcgacagt aac**g**tga cgaacgcaatgccaagagatctgac
core5	cgttagacagtgccgcatcgacagt aac**t**tga cgaacgcaatgccaagagatctgac
core6	cgttagacagtgccgcatcgacagt aacc**a**ga cgaacgcaatgccaagagatctgac
core7	cgttagacagtgccgcatcgacagt aacct**a**a cgaacgcaatgccaagagatctgac
core8	cgttagacagtgccgcatcgacagt aacctg**g** cgaacgcaatgccaagagatctgac
opti9	cgttagacagtgccgcatcgacagt aacctg**c** cgaacgcaatgccaagagatctgac
opti10	cgttagacagtgccgcatcgacagt **c**acctga cgaacgcaatgccaagagatctgac
opti11	cgttagacagtgccgcatcgacagt aac**a**tga cgaacgcaatgccaagagatctgac
opti12	cgttagacagtgccgcatcgacagt aacc**c**ga cgaacgcaatgccaagagatctgac
opti13	cgttagacagtgccgcatcgacagt aacc**g**ga cgaacgcaatgccaagagatctgac

Table 2.4 Sequence comparison of the different Six4bss variants (designated Six4bssmut). Each sequence is removed from the SELEX derived Six4bss by a single nucleotide substitution. Bold lettering indicates deviations from the experimentally determined consensus sequence.

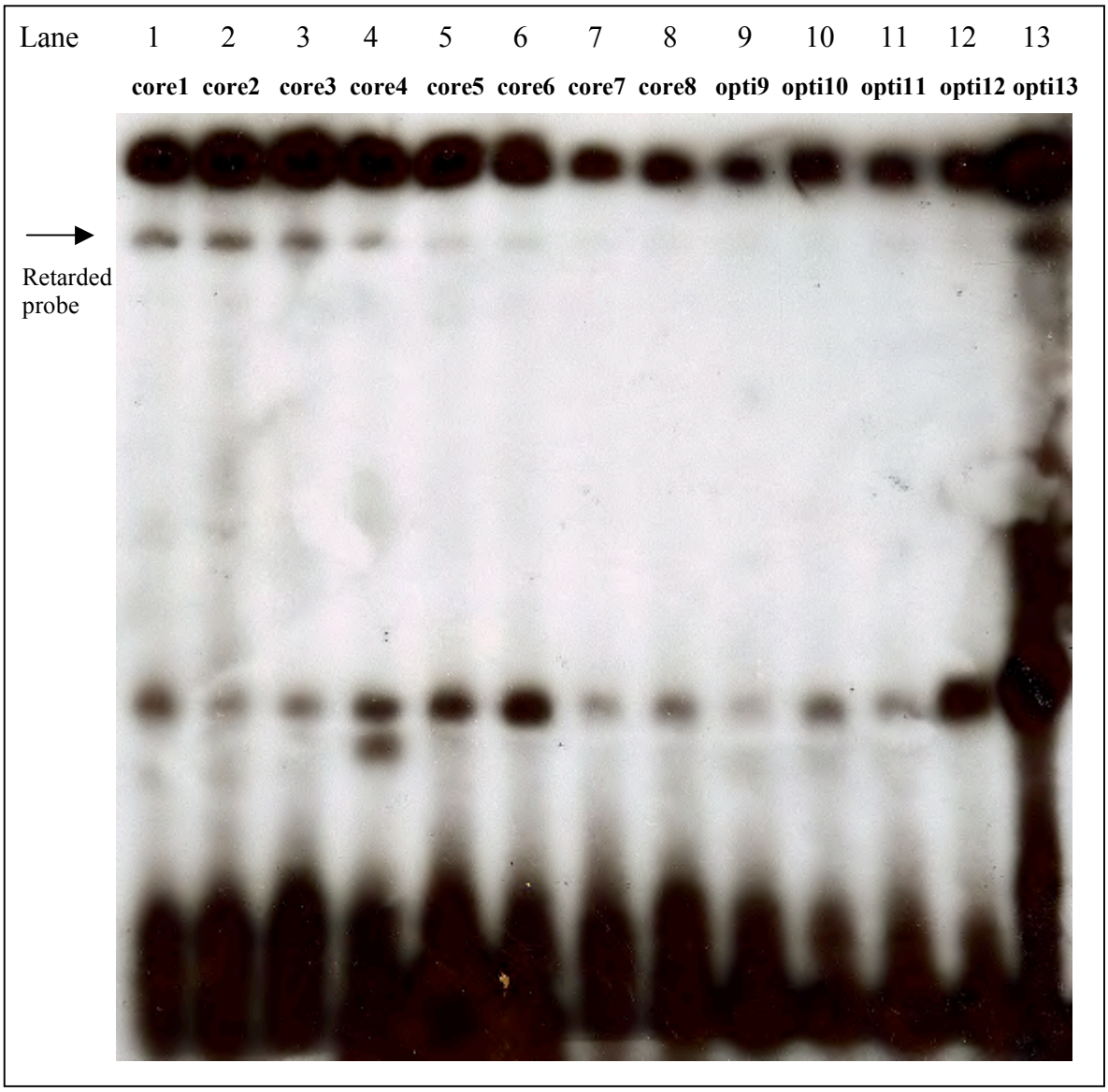


Fig. 2.17 Relative in vitro binding of different Six4bss variants by W.C.Es of cells containing the GST-SD+HD+ encoding plasmid. In a 30-ml binding reaction, a W.C.E extract was incubated with a radiolabelled Six4bssmut probe at room temperature for 1 h. Binding reactions were subsequently loaded into an 8%polyacrylamide gel in Tris-borate buffer. The relative intensities reflect the relative affinities of the proteins for the DNAs. Lanes 1-13 contain contain radiolabelled oligos with a designation matching the lane number (For complete sequences see table 2.4). Black arrow indicates shifted protein-DNA complexes.

2.20 Generation of a refined Six4bss consensus sequence

In light of the findings of the Binding Sequence Position Occupancy analysis the Six4bss multi-level consensus sequence was redefined so as to incorporate the new findings. As discussed above however, this new consensus sequence does not have the benefit of a positional weight matrix defining the likelihood of incorporation of different nucleotides at different positions. This is because the consensus was generated through 2 different methods (SELEX and EMSA) and therefore the likelihood of incorporation of each base at each position is deemed to be equal despite previous observations. The utility of this matrix is discussed later. A weighted consensus sequence was derived in a semi arbitrary fashion. The position occupancy likelihoods derived from the original Six4bss were supplemented with the findings of the EMSA analysis by adding consensus conforming sequences. For a representation of the revised Six4 consensus binding sequence see table 2.5 along with position occupancy information and comparisons to previously reported Six5 binding sites see table 2.5.

Sequence									
Six4bss permitted	G	T	A/C/G	A/G	C/T	N	N	G/A	A/G
Six4bss consensus	G	T	A	A	C	C	T	G	A
MEF3 (reverse)			A	A	C	C	T	G	A
TREX (permitted)		C/A/T	A	C/T	C/T	C/T/G	G	A	G/A/T
ARE	G	C	A	A	C	C	T	G	A CACC
***MCK* TREX**		C	A	C	C	C	G	A	G

Table 2.5 Comparison between the Trex permitted sequence (based on individual base pair mutation, see sections 2.19 and 2.20), the Trex sequence from *MCK* enhancer, the *ARE* enhancer, the MEF3 enhancer, the SELEX-derived Six4bss consensus and the permitted Six4bss (based on individual base pair mutation)

2.21 Discussion

The target detection assay performed using the recombinant GST-SD+HD protein has led to the identification of a putative Six4 consensus binding sequence. This sequence was found to be highly similar to that recognised by other Six4/5 subfamily members. Moreover a positional weight matrix for the Six4 binding sequence was constructed and the mechanics of a SELEX have been investigated and refined. What follows is an analysis of the significance of these findings as well as a number of considerations that need to be made in their interpretation.

2.21.1 *In vitro* binding sequence determination limitations

This analysis is a completely *in vitro* approach and has elucidated the ligand-DNA binding properties of the Six4 homeo- and SIX domains. In reality the *in vivo* DNA binding specificity of transcription factors can be influenced by factors that can't be represented in an *in vitro* reaction such as the presence of cofactors that can bind to transcription factors and alter or define their binding specificities (Murre et al., 1989). As has been previously mentioned SIX4 and SIX5 are shown to form a functional heterodimer with Eya proteins (Ohto et al., 1999) and Six4 is strongly suspected of interacting with the *Drosophila* Eya homologue (Ivan Clark, personal communication). It is however known that SIX5 *in vitro* targets isolated through SELEX match those shown *in vivo* (Rami Jarjour, personal communication) and based on sequence comparisons of both the SIX and homeodomain of both Six4 and SIX5 (Kirby et al., 2001) it would be reasonable to assume that Six4 would have a similarly unaltered binding specificity *in vitro*. It is therefore likely that a genome search conducted using the knowledge on the Six4 binding specificity would identify potential Six4 regulation targets.

Additionally a comparison between the homeo- and SIX domains of Six4 to mouse and human Six4 and Six5 reveals a conservation level of 59% and 56% respectively in the case of the SIX domain and, more importantly, 81% with 83% in the case of the homeodomain (Figures 2.21 and 2.22). Most substitutions within the homeodomain occur in positions that have been shown to be under weak evolutionary constraints between most of the known SIX4/5 homologues and are thus likely to be of little importance in conferring DNA binding specificity (Raphaela Kitson-Pantano, personal communication). Moreover these positions are often different between human and mouse Six4 and Six5 genes. Given the observed redundancy between

these genes and the observed unaltered binding properties of the *Drosophila* and mouse SIX4/5 homologues it is likely that these positions are of little importance in conferring DNA binding specificity (Figures 2.21 and 2.22). This corroborated the theory that the DNA binding specificity of the SIX4/5 subfamily remains largely unchanged throughout vertebrate evolution and inferences made for SIX5 regulatory targets in mammals may hold true for their *Drosophila* homologues.

Additionally, given the nature of developmental network evolution, it is unlikely for the binding specificity of a transcription factor of apparent multisystemic utility such as SIX4/5 to adopt any appreciable alterations in the face of such conservational pressure. However, it is possible for minute changes in specificity to occur since the duplication of the SIX4/5 ancestor gene as selective pressures are relaxed although the apparent redundancy between SIX4 and SIX5 suggests that this is not the case.

```
dSix4        1   STDQIQCMCEALQQKGDIEKLTTFLCSLPPSEFFKTNESVLRARAMVAYNLGQFHELYNL
mSix4        1   -P-HVA-V------G-NLDR-AR--W---Q-DLLRG---L-K---L--FHQ-IYP---SI
hSix4        1   -P-HVA-V------G-NLDR-AR--W---Q-DLLRG---L-K---L--FHQ-IYP---SI
mSix5        1   -PE-VA-V----L-A-HAGR-SR--GA---A-RLRGSDP------L--FQR-EYA---Q-
hSix5        1   -PE-VA-V----L-A-HAGR-SR--GA---A-RLRGSDP------L--FQR-EYA---R-

dSix4       61    LETHCFSIKYHVDLQNLWFKAHYKEAEKVRGRPLGAVDKYRLRKKYPLPKTIWDGE
mSix4       61    --S-S-ESAN-PL--Q--Y--R-T---RA--------------R-F---R------
hSix4       61    --S-S-ESAN-PL--Q--Y--R-T---RA--------------R-F---R------
mSix5       61    --SRP-PAAH-AF--D-YLR-R-H---RA---A------------F----------
hSix5       61    --SRP-PAAH-AF--D-YLR-R-H---RA---A------------F----------
```

Fig. 2.21 Sequence alignment of the protein sequences of the SIX domains of *Drosophila* Six4 (dSix4), murine Six4 (mSix4), murine Six5 (mSix5), human SIX4 (hSix4) and human SIX5 (hSix5). The (-) symbol indicates an invariable position.

```
dSix4        1  WDGEETVYCFKEKSRNALKDCYLTNRYPTPDEKKTLAKKTGLTLTQVSNWFKNRRQRDRTP
mSix4        1  -------------------EL-KQ----S-A--RH---I---S----------------N-
hSix4        1  -------------------EL-KQ----S-A--RH---I---S----------------N-
mSix5        1  ------------R--A---A--RG---------RR--TL---S------------------
hSix5        1  ------------R--A---A--RG---------RR--TL---S------------------
```

Fig. 2.22 Sequence alignment of the protein sequences of the homeodomains of *Drosophila* Six4 (dSix4), murine Six4 (mSix4), murine Six5 (mSix5), human SIX4 (hSix4) and human SIX5 (hSix5). The (-) symbol indicates an invariable position.

2.21.2 Significance of DNA-protein interactions detected *in vitro*

An additional issue that needs to be addressed is whether all the DNA-protein interactions that can be detected *in vitro* actually have implications for *in vivo* applications. Just because a protein is bound specifically at its DNA target site, does that fact alone indicate a *cis*-regulatory function? There are thermodynamic and probabilistic arguments which point toward the functional significance of many specific DNA-protein interactions**.** Such arguments include the fact that *in vitro* determined binding sequences are ultimately obeying the same rules for ligand-DNA binding as their *in vivo* counterparts (a fact that has been confirmed by X-ray crystal structures of many such complexes). Arnone and Davidson (1997) state that ‘‘the regulatory significance of transcription factor- DNA interactions is demonstrable experimentally in gene transfer experiments and functions have been identified by this means for certain transcription factors’’ Most of these issues have been addressed in the above sections but ultimately it is difficult to definitively claim that an *in vitro* interaction is indicative of the behaviour of a ligand *in vivo*. However it is my belief that the independent identification of a binding sequence for Six4 that bears an uncanny resemblance to the Six5 *in vivo* binding site is a very strong indication that this finding is important.

Additional information concerning the binding specificity of Six4 could potentially be obtained through the use of by NMR spectroscopy to study the complex of the SIX and homeodomains with a DNA fragment corresponding to the Six4 binding site in aqueous solution. This method would determine the structure of the Six4 homeodomain and allow for inferences on positional base occupancy to be made.

2.21.3 On the use of a derived positional weight matrix in identifying putative Six4 regulatory targets

Even in the absence of entirely exhaustive knowledge on the DNA-binding specificity of a transcription factor, the possession of a high confidence binding site recognition tool such a detailed Positional Weight Matrix (PWM) can provide useful information on direct interactions between participants in developmental pathways such as Six4. The next chapter proceeds to utilise the information gleaned from this SELEX analysis to query the *Drosophila* genome for putative downstream targets of Six4.

2.21.4 Possibility of multiple DNA binding specificity

This study has chosen to disregard the presence of isolates that have computationally been determined to not contribute to the consensus sequence or bear a minimal resemblance to previously reported SIX4/5 subfamily binding sequences. No discernible similarity beyond that which can normally be expected for random sequences of a given length was observed between those isolates. However, the fact that these dissimilar isolates comprised ~ 60% of all sampled sequences could suggest a role for these isolates beyond that of being by-products of an imperfect selective process. It is conceivable that the inclusion of non-consensus conforming isolates is the result of the ability of Six4 to bind to 2 or more different target sequences. There is precedent to these occurrences, such as the ability of the TEF-1 factor to recognise different sequences (Yoshida, 2008). It is indeed likely that this interchangeable specificity contributes to regulation of gene transcription in more context specific manner and may rely on the presence of different co-factors. There has been no indication that Six4 acts in this way however, and this study will choose to ignore the possibility of Six4 having two different binding sequences.

2.21.5 On the implications of the interspecific conservation of SIX4/5 subfamily binding sequences

As mentioned previously, the SELEX determined binding sequence of Six4 is highly similar (though not identical) to that recognised by the murine Six5. This fact is not in itself surprising given the conservation of the homeodomains involved in these interactions. This conservation is however indicative of a possible higher degree of conservation that could characterise the entire developmental pathways surrounding the SIX proteins.

APPENDIX 2.1:

10 mM Tris-HCl	$MgCl_2$	25 mM KCl	75 mM KCl
pH 8.3	1.5 mM	Opti-Prime™ 1× buffer #1	Opti-Prime™ 1× buffer #2
pH 8.3	3.5 mM	Opti-Prime™ 1× buffer #3	Opti-Prime™ 1× buffer #4
pH 8.8	1.5 mM	Opti-Prime™ 1× buffer #5	Opti-Prime™ 1× buffer #6
pH 8.8	3.5 mM	Opti-Prime™ 1× buffer #7	Opti-Prime™ 1× buffer #8
pH 9.2	1.5 mM	Opti-Prime™ 1× buffer #9	Opti-Prime™ 1× buffer #10
pH 9.2	3.5 mM	Opti-Prime™ 1× buffer #11	Opti-Prime™ 1× buffer #12

The final reaction concentrations of the various buffer components utilised in PCR optimisation described in section 2.11 are listed in the top row: the first column depicts the buffer pH of Tris-HCl, the second column depicts the $MgCl_2$ concentration and the last two columns depict the KCl concentration of each buffer (The above table can be found in the Opti-Prime™ Kit user's manual, Stratagene).

Chapter 3 – Identification of putative targets of Six4 regulation

3.1 Introduction

One of the goals of this study is to use the information generated through SELEX and described in Chapter 2 to identify putative transcription factor binding sites (TFBSs) for Six4. Various approaches have been utilised so far to that end. This chapter will discuss the use of a number of bioinformatics approaches for the discovery of TFBSs. These methods are probabilistic and cannot match the accuracy of, or indeed completely substitute 'wet' experimental data. They are however more cost-effective and have the potential to generate information much faster. All of these methods essentially involve the creation of a search tool from a multiple alignment of the sequences of the isolated aptamers through the use of a classification algorithm.

TFBS classification algorithms generate a numerical score representing the degree to which a given sequence site matches a given motif. Most of these algorithms utilise a scoring model that can take many forms, such as a fixed-order Markov model or simply a position weight matrix (PWM), or a Hidden Markov Model (HMM).

The purpose of this study was to test both the stringency and sensitivity of some of these approaches, and respective algorithms, in detecting putative Six4 binding sites based on the information generated through SELEX.

3.2 Positional Weight Matrices

The most prevalent context-independent model for detection of TFBSs is the Position Weight Matrix (PWM) (or position specific score matrix—PSSM). Simple PWMs are widely used in conjunction with public access transcription factor databases (like TRANSFAC^{TM}) for the identification of TFBSs. They are essentially a representation of the observed frequencies of the nucleotides at each position of an alignment present in a training set. These frequencies translate into scores or ''weights'' assigned to different expressions of a position in a screened sequence. The individual weights of the different positions are combined to generate an overall weight based on which the decision of whether or not the sequence being screened constitutes a putative TFBS is made. PWM models (and the algorithms based on them) often have no context dependencies at all (Match^{TM}, Kel et al., 2003). In that respect they can be considered to be fixed-order Markov models of order 0, also known as Bernoulli models (each site is assumed to evolve within its own constraints and tendencies, free of the influence of any neighbouring sites). The major pitfall of this assumption is that it presupposes that states at different positions are statistically

independent. Thus essentially ‘‘the joint probability of finding a multiple-position site factorizes into the product of single-position probabilities’’ (Ben-Gal et al., 2005). However the dependence between positions in TFBSs is a documented fact (Bulyk et al., 2002). Therefore the independence assumption made by most PWM models is violated in the mechanics of most TFBSs. In the particular case of TFBSs this assumption is not nearly as fundamentally flawed as it would be in the case of RNA where factors like secondary structure come into play. The influence exerted by TFBS position interdependencies is often overridden by the overall efficiency of some PWMs. Therefore, in spite of this shortcoming, the overall performance of some PWMs can be satisfactory, hence their popularity. Ben-Gal et al. (2005) state that ‘‘the PWM model, which is based on the (unsupported) independence assumption, is often found to outperform fixed-order Markov models of higher order that are based on the (reasonable and supported) dependence assumption’’. Additionally Posch et al. (2007) state that ‘‘PWM models may outperform Markov models of higher order’’ because the limited amount of experimentally verified binding sites available for the learning phase may result in the problem of overfitting for models with larger numbers of parameters (like Hidden Markov Models, see below).

Open access algorithms that generate and/or use PWMs include PATCH™ (Kel et al., 2005), Profilemake (Gribskov and Veretnik, 1996), MatInspector™ (Cartharius et al., 2005) and TESS™ (Schug and Overton, 1997). Most of these algorithms are based on the same overarching concept and a good knowledge of the parameters involved in performing an analysis can often render them virtually interchangeable. The performance of one such algorithm at detecting Six4 binding sites will be tested in this study (matrix-scan).

3.3 Hidden Markov Models

Another approach towards detecting TFBSs involves the use of a Hidden markov Model (HMM). A profile HMM is a statistical model of multiple sequence alignments. It essentially represents position-specific information about how conserved each column of an alignment is, and which residues are likely to occupy which positions, much like a PWM (of which it is essentially a more parameterised version). HMMs determine the hidden parameters (in this case position weights and interdependencies) from the observable parameters (the position states in the initial alignment or training set). The extracted model parameters can then be used to perform further analysis such as pattern recognition. ‘‘All of the profile methods are

more or less statistical descriptions of the consensus of a multiple sequence alignment.''(Sean Eddy in HMMER2 user's guide Version 2.3.2; Oct 2003). As mentioned previously Markov models are characterised by an order. A Markov model of order *m* means that the probability of each residue depends on the *m* preceding residues in the sequence. The order of most PWMs used by available algorithms is 0, whereas the order of an HMM built for the same purpose depends on the sequences populating a training set.

HMMs have a formal probabilistic basis. Probability theory is used to determine the scoring parameters. The use of an HMM removes all bias (with the exception of generating the multiple sequence alignment) from a whole genome screen and incorporates more of the sequence binding information generated by the SELEX approach than a PWM. Essentially, when built on a large enough (and representative) training set an HMM incorporates more of the information concerning position occupancy contained within that set. What the user must however understand is that the use of an HMM also opens the door for bias in pattern recognition resulting from a skewed training set.

An evaluation of methods for TFBS detection involving the use of a HMM constructed using the HMMER utility (see below) (Eddy, 1998) showed that an HMM outperformed four other widely used tools both in identifying seeded TFBSs in an experimental training set but also generated less false positive matches (Marinescu et al., 2005). The utilities evaluated in that study included ''Match and Patser that scan a sequence using a supplied PWM, LMM (Local Markov Method, see below) that uses a p-value-based scoring measuring the similarity of the hit to the known binding sites for the factor and its contrast to the local genomic context, and ScanACE that scans a sequence for matches for a given motif using a scoring method based on a maximum *a priori* log likelihood score''(Marinescu et al., 2005). The main advantage a HMM has over a PWM is the fact that it makes no assumption during model generation and therefore does not violate any biological axioms (by assuming positional independence of TFBSs). However, a potentially crucial consideration is that HMMs have been reported as being uninformative if they are based on short sequence alignments (like those involved in TFBS specification) because of the small amount of data that is available for training the model (Mount, 2004). A number of interfaces for the construction and use of HMMs in the identification of TFBSs are publicly available (MAPPER, Marinescu et al., 2005, HMMER at http://bioweb2.pasteur.fr;

Eddy, 1998). Essentially, most make use of the HMMER algorithm and accompanying group of programs.

The cardinal difference between the two approaches is the number of parameters involved. A PWM is heavily based on the, incorrect, independence assumption, and thus results in an under-fitted model with a smaller-than-necessary number of parameters. On the contrary, HMMs have a large dimensionality but when generated from a small training set tend to be over-fitted. That means that a PWM will outperform (see below for performance evaluation) an HMM when the latter is modelled on a small training set. What constitutes a small training set is a matter of debate and there is no right number of sequences that will constitute a comprehensive training set. Model performance depends entirely on training set composition. Hannenhalli (2008) states that ‘‘PWM representation assumes independence among positions within a binding site, a full dependence model, on the other extreme, requires estimating an exponentially large joint distribution based on a small number of exemplars. The optimal choice among these possibilities may vary among TFs and a detailed evaluation of these choices needs to be done.’’ This study will therefore conduct an independent evaluation of available algorithms that use these two approaches in an attempt to define a robust method for detecting TFBSs using the available SELEX data. The algorithms utilised in this study are Matrix-Scan (developed by Jean Valéry Turatsinze, Morgane Thomas-Chollier and Jacques van Helden, available at http://rsat.scmbb.ulb.ac.be/rsat/, Thomas-Chollier et al., 2008), and HMMER2 (release 2.3.2, Eddy, 1998) for HMMs. These approaches come with their own supporters and detractors but essentially their suitability to the task at hand depends on the particulars of that task. These particulars include the nature of the TFBS being sought and the available collection of binding sequences that constitutes a representation thereof (SELEX aptamer collection). With these facts in mind this study has performed an unbiased comparison of the performance of these approaches using a wide range of parameters and arrived at a preferred method and parameter set for screening the *D.melanogaster* genome for putative TFBSs of Six4.

It is worth mentioning before continuing that other methods for the identification of TFBSs are available that combine the properties of PWMs and fixed order markov models (like HMMs). Such models include variable order Bayesian networks (VOBN, Ben-Gal et al., 2005) that do not make the assumption of position independence but only take into account position interdependencies that are statistically significant. These approaches were not considered due to the lack of availability of their

associated algorithms (VOBN and VOMBAT, Ben-Gal et al., 2005; Posch et al., 2007). Finally, many of the methods described in this chapter have been successfully used in *Drosophila* to identify direct regulatory targets of the retinal determination protein Eyeless (ey)(Ostrin et al., 2006), a known regulator of the SIX protein Optix.

3.4 Generating a multiple sequence alignment (MSA) from the selected SELEX aptamers

Essentially, a TFBSs detection model will only ever be as informative as the alignment it is based upon. All the information available to this study concerning the position occupancy of the Six4 binding sequence is contained within the sequences of the isolated SELEX aptamers, as well the results of the TFBS single position mutagenesis (see chapter 2). Given the apparent heterogeneity evidenced in the SELEX sequence set it was deemed necessary to isolate a subset from within the pool of selected isolates the members of which are very similar. Otherwise any resulting MSA would not have a high enough information content to generate a useful model. The information content of a model essentially represents its ability to distinguish a genuine hit from background. The concept of information content is a well established one and the content itself is measured in bits (the maximum absolute information content at each position is 2, this signifies an invariable position, this concept is explained in greater detail in section 3.6.2). In order to obtain the alignment with the highest information content, this study performed a profile analysis of the SELEX results. This essentially involved performing a global MSA of all the resulting aptamers and removing the most highly conserved aptamers into a smaller MSA. This MSA was then used to inform the resulting PSSM or HMM. The disadvantage of this approach is that the resulting models will only be as representative of the variation in position occupancy as the training MSA itself. It is likely that there is genuine information inherent in the non MSA conforming aptamers that is therefore omitted from the resulting model. However given the difference in the information content between the global and local MSAs this risk was deemed worth taking. Essentially the information content of an MSA is directly linked to the information content of any resulting matrices and/or model (see below).

The concept of MSA refinement is recurring one and numerous available algorithms have been designed to tackle this problem (see algorithm evaluation by Chakrabarti et al., 2006). Most of these algorithms however deal with protein MSA

refinement. Luckily this study need not concern itself with the complexities of MSA generation that plague protein sequence alignment such as gap introduction and spatial position calculation. The TFBSs present within the SELEX aptamer collection are short uninterrupted sequences of roughly the same length. As such the problem of selecting a subset of sequences that will generate a high information model is one of Motif elicitation rather than MSA refinement. It is a question of identifying a recurring pattern from within a sequence set rather than trying to align all the available sequences. The motif elicitation algorithm used by this study was MEME™ (Bailey and Elcan, 1994). MEME is perhaps the most common used algorithm in motif elicitation and provides dependable results when searching for short gapless motifs. Other algorithms used for this purpose include GLAM (Frith et al., 2004, http://zlab.bu.edu/glam/), motifSampler and alignACE. Fu and Weng (2004) report that GLAM actually outperformed MEME, motifSampler and alignACE (Roth et al., 1998, http://atlas.med.harvard.edu/) in identifying motifs for TFBS PWM generation. A-GLAM (a variant of GLAM used for detecting gapless motifs) was also used in this study and the same motif as the one reported by MEME was reported.

An initial analysis of the 39 available isolates (only the variable core sequences, sequences can be seen in Fig. 3.1) by MEME (default parameters, requested motif size took all values between 6-13 based on previous SIX family binding site observations) revealed a 7bp long motif present in 16 sequences (Fig. 3.1) The information content of the PSSM resulting from the MSA of all the motif containing sequences was shown to be 10.5 bits based on the observed nucleotide frequencies (out of a possible 14, also see section 3.6.2).

A curious observation was made when aligning the above sequences. Most of the aligned sequences also included the nucleotides G and usually T directly upstream of the highlighted motif. The inclusion of these two positions in the identified motif would generate a consensus alignment that was determined to be GTAACCyGA. However, some of the sequences were shorter (by virtue of being products of the R57 pool of oligos that contained a 7 bp random core) and thus did not align over the full length of their sequence. This suggested that the excluded nucleotides in these cases were either not required for binding or that the necessary nucleotides were provided by the flanking non-random arms of the oligo. It was often the case that the addition of 2-3 nucleotides from the primer-annealing arms to one of the ends of the sequence completed the alignment. This was done in a careful manner so as not to shift the focus of the alignment onto the flanking sequence. The added sequence consisted of a

total of 12 nucleotides distributed over 6 sequences. The added nucleotides, where present are included in black boxes in Fig. 3.4. The SELEX sequences along with the added nucleotides were analysed by MEME and a 9bp long motif present in 17 sequences was identified. The resulting PSSM had an information content of 13.4 (~1.49 bits per position) compared to 10.5 (~1.5 bits per position) for the initial PSSM. A higher MSA information content is preferable when generating a matrix since it enhances the selectivity and specificity of the model.

Since the flanking arms were part of the aptamers, and therefore contributed to their ligand-binding affinity, the inclusion of those nucleotides did not violate the principles of SELEX and helped to generate a more robust model for the detection of Six4 bss. Interestingly, the addition of these extra nucleotides enhances the performance of most models generated by a MSA that includes them, both in terms of specificity and selectivity (see below) as evidenced by both the increased computed information content as well as the data presented herein. For the purposes of this comparison the alignment that excludes these nucleotides will be termed *align1* and the alignment that includes them will be termed *align2*. The final alignment can be seen in Fig. 3.4 and constitutes the basis of all the models utilised in this study.

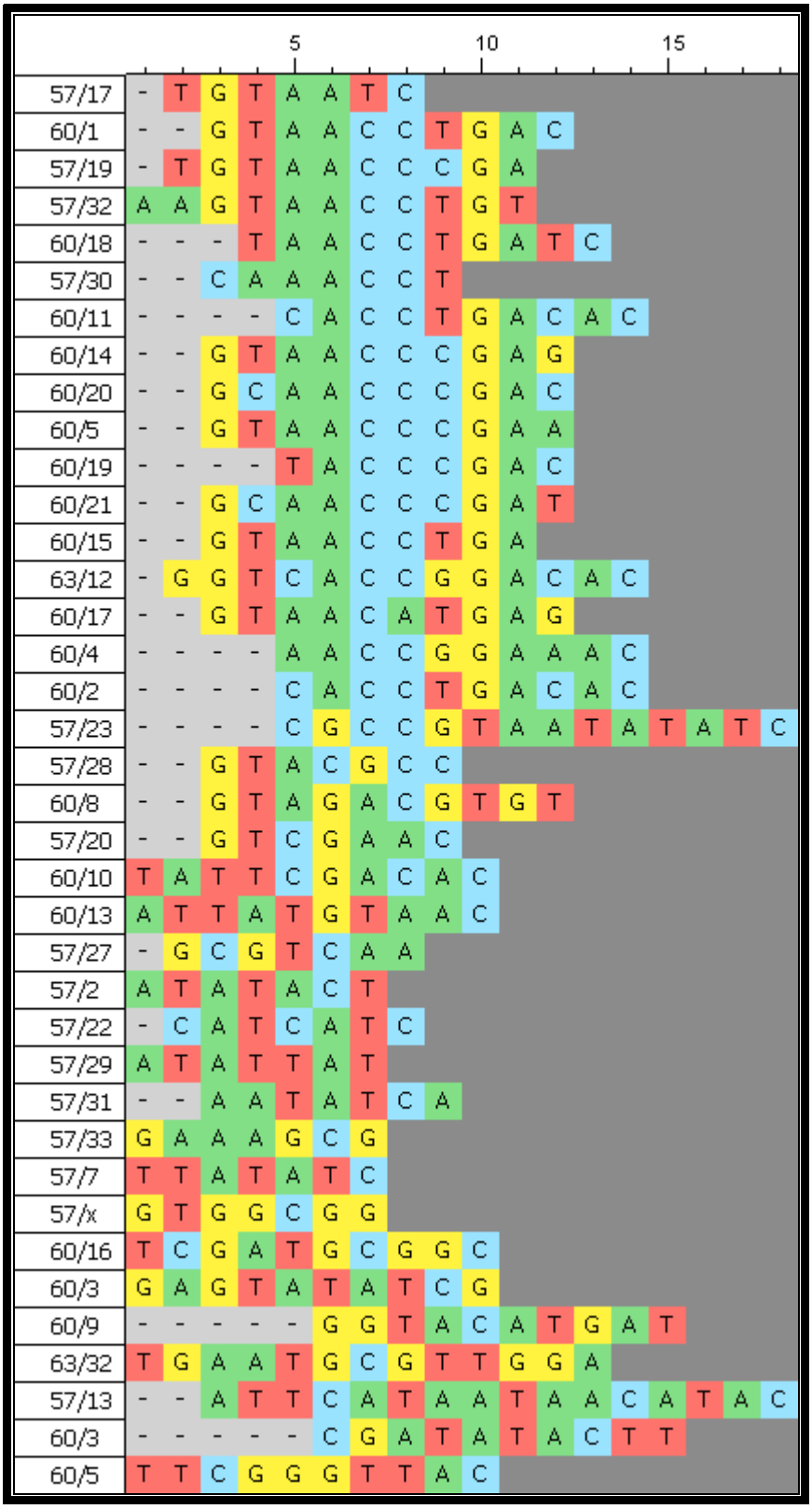


Fig. 3.1 Alignment of all the variable regions of all the aptamers isolated through SELEX. For the purposes of generating this alignment, sequences were mostly aligned manually using the CINEMA (Colour INteractive Editor for Multiple Alignments) utility contained in the UTOPIA program pack (Pettifer et al., 2004)

NAME	STRAND	START	P-VALUE	SITES
60/15	+	3	5.91e-05	GT AACCTGA
60/18	+	2	5.91e-05	T AACCTGA TC
60/1	+	3	5.91e-05	GT AACCTGA C
60/21	+	3	1.22e-04	GC AACCCGA T
60/5	+	3	1.22e-04	GT AACCCGA A
60/20	+	3	1.22e-04	GC AACCCGA C
60/14	+	3	1.22e-04	GT AACCCGA G
57/19	+	4	1.22e-04	TGT AACCCGA
60/2	+	1	2.47e-04	CACCTGA CAC
60/4	+	1	2.47e-04	AACCGGA AAC
60/11	+	1	2.47e-04	CACCTGA CAC
60/17	+	3	4.28e-04	GT AACATGA G
60/19	+	1	4.91e-04	TACCCGA C
57/32	+	5	6.09e-04	AAGT AACCTGT
63/12	+	4	6.76e-04	GGT CACCGGA CAC
60/9	+	3	1.04e-03	GG TACATGA T

Fig. 3.2 The 7-bp long motif, initially identified by MEME in 14/41 sequences. Resulting MSA has an information content of 10.5. Aptamer core sequences and designations are provided. Per base information content is normalised based on the nucleotide frequencies observed in the starting sequence set.

NAME	STRAND	START	P-VALUE	SITES
60/15	+	1	3.62e-06	GTAACCTGA
60/18	+	1	3.62e-06	GTAACCTGA TC
60/1	+	1	3.62e-06	GTAACCTGA C
60/5	+	1	7.57e-06	GTAACCCGA A
60/14	+	1	7.57e-06	GTAACCCGA G
57/19	+	3	7.57e-06	-T GTAACCCGA
60/2	+	1	1.55e-05	GTCACCTGA CAC
60/4	+	1	1.55e-05	GTAACCGGA AAC
60/11	+	1	1.55e-05	GTCACCTGA CAC
60/17	+	1	2.67e-05	GTAACATGA G
60/19	+	1	3.46e-05	GTTACCCGA C
60/21	+	1	4.26e-05	GCAACCCGA T
60/20	+	1	4.26e-05	GCAACCCGA C
57/32	+	3	5.34e-05	AA GTAACCTGT
57/17	+	2	5.71e-05	T GTAATCCGA
63/12	+	2	6.96e-05	G GTCACCGGA CAC
60/9	+	1	3.69e-04	GGTACATGA T

Fig. 3.3 Revised 9-bp long motif generated after the addition of 12 nucleotides present in the constant flanking arms of selected aptamers. Resulting MSA has an information content of 13.4 bits. Aptamer core sequences and designations are provided. Per base information content is normalised based on the nucleotide frequencies observed in the starting sequence set.

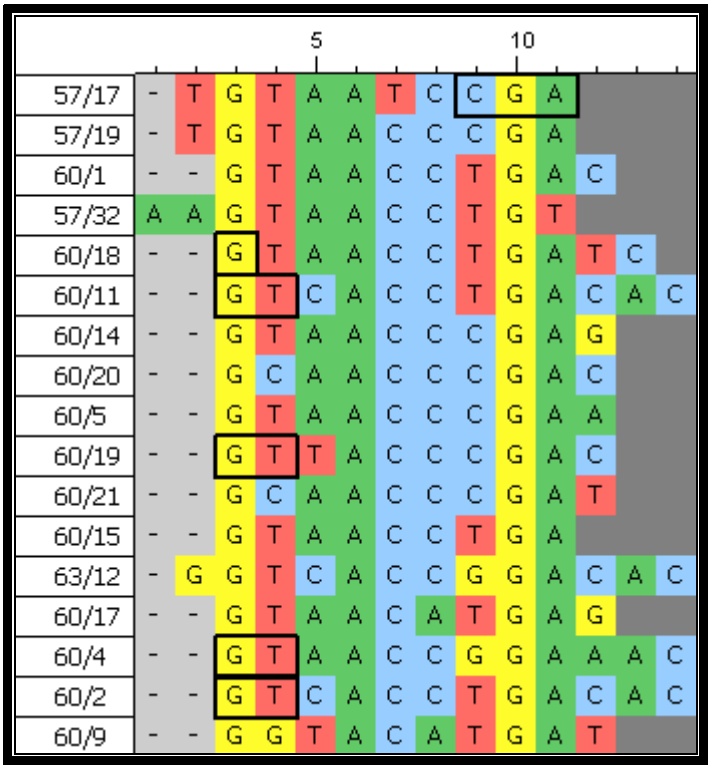


Fig. 3.4 Multiple sequence alignment of the 17 isolates used for putative Six4 binding sequences (*align2*). The boxed nucleotides indicate additions from the non-random flanking arms of the aptamers. Graphical Alignment generated through CINEMA^{TM} (Pettifer et al., 2004).

3.5 Assessment of Model performance

The purpose of a classifier system (model) is to identify potentially meaningful sequences (putative TFBSs) from within a sequence library. This section will introduce a number of concepts that will be used in the evaluation of different models. Most of these concepts are context-sensitive and their definition relies upon the experimenter's appreciation of them. To a classifier system a true positive (TP) is a hit that is considered to be meaningful because of its functional significance (biological or otherwise). In terms of transcriptional regulation, true positives would be putative TFBSs detected by a model that bind transcription factors in *vivo*. Conversely, a false positive (FP) hit is a sequence recognised by a model that is of no importance in TF binding. The nature of sequence modelling is such that the distinction between the two is almost always beyond the ability of sampling algorithms to make. Therefore additional testing is required to tell the difference between the two. This issue is dealt with is sections 3.6-3.8. Finally, the concepts of true and false negatives (TN and FN), although important to most classifier systems, are of less concern to this study since most models deal with ''hits'' (reported positives) rather than ''misses'' (everything else in a sampled sequence). Also, the number of FN is intrinsically linked to the TP and is as such indirectly addressed in the following tests. Henceforth, when the terms TP, FP, TN and FN are mentioned they will refer to the numbers of such hits or misses detected in a sample.

In order to establish the best approach for detecting putative Six4 binding sites I assessed the performance of both a PWM as well as a number of HMMs on the basis of two criteria. i) Sensitivity (or recall) i.e. the ability of a model to detect genuine Six4 bss or ''true positives'' expressed as the true positive detection rate [TP / (TP + FN)] and ii) the positive predictive value, or precision rate i.e. the ability of a model to distinguish between true and false positives and minimise the occurrence of the latter [TP / (TP + FP)]. In order to assess model performance a set of test sequences was generated and used to measure these values for all the generated models.

The models were tested by embedding the sequences for 10 of the 17 isolates (Fig. 3.5, associated weight scores for individual sequences are provided based on a PWM generated from *align2*) that contributed to the model generation in a random genomic sample that consisted of 500 randomly generated 2kb-long sequences. This was done to assess for the ability of the model to detect genuine Six4 binding sites and distinguish them from genomic background. These random sequences were obtained

through the random sequence retrieval utility at RSAT (Regulatory sequence analysis tools, http://rsat.scmbb.ulb.ac.be/rsat/, random seq, 1997- 2007, Jacques van Helden, 2003; Thomas-Chollier et al., 2008) using a *Drosophila melanogaster*-upstream region specific Markov chain (Cuticchia et al., 1992) (essentially a model that generates sequences based on observed base transition frequencies observed in *Drosophila melanogaster* upstream regions). The generated DNA sequences were calibrated on non-coding upstream sequences since this is where TFBSs are likely to be located. This should in principle allow for assessment of the performance of generated models in detecting Six4 TFBSs in native *Drosophila* sequences. The test sequence sets were designated *random1*, for the unaltered 2000 random sequence collection and *random2,* for the sequence library containing the 10 ‘‘planted’’ aptamer target sequences. The presence of the 10 planted sequences at known locations within the sequence library is the only difference between *random1* and *random2*. It is worth mentioning that the 10 planted sequences included the 3 sequences that generated the lowest scores using the *align2* PWM (see section 3.6, Table 3.1, sequences 4, 5 and 8). This was done so as to assess the ability of the generated models to detect ‘‘weak’’ members of the alignment. Both random sequence sets were then probed with the constructed HMMs as well as the Six4 PWM (see below) using a variety of different parameters and assessed for sensitivity and precision rate. Given the difference in the parameters that need to be evaluated under the two methods, the results are presented in different sections.

The weights assigned to the planted sequences by the PWM are included in Table 3.1. Weights were generated using the PWM shown in Fig. 3.7. For HMMs, assigned weights are less meaningful since they differ based on the utilized null model (see section 3.7). In those cases evaluation was performed as detailed in section 3.7. Knowledge of the assigned weights allowed for making an informed decision when setting a discrimination threshold (cut-off point).

It is worth mentioning that this analysis does not directly allow for the assessment of the number of false positives one expects to find in native genomic sequences but it allows for the direct comparison of different models and parameter sets in terms of sensitivity and specificity (or in this case precision rate). This is because even the best model can not distinguish between a genuine binding site and a biologically insignificant one if they both have the same sequence. This is a relative test designed to assess the performance of the two different methods and as such makes a number of arbitrary assumptions. The test initially assumed that only the planted sequences can

generate true hits. During the test other sequences that were often identical to the planted ones were detected within the randomly generated sequence sets. Hits generated from those were considered to be false (this is an arbitrary classification that is acceptable for comparative purposes only).

In more realistic terms, the ability of a classifier to detect planted sequences that are identical to the ones used in the training set can be seen as a measure of its efficiency. However, there must always be a finite number of 'true' positives if one is to calculate the sensitivity value. If one is willing to call all sequences that match the training set 'true' positives then the willingness to use a weight-based classifier comes into question. That is because such a classifier is by definition used to identify sequences that may not be included in the training set. If only the training set sequences are to be assigned significance then a sequence search (a much more simplistic approach that only detects exact matches to a sequence) should be used. However, for reasons mentioned previously, this study is willing to assume that the binding specificity of Six4 may encompass (and indeed does as shown by the binding site position occupancy analysis experiment described in section 3.19) sequences not isolated through SELEX (and as such absent from the training set). Therefore, assigning significance only to exact matches to the training set violates the assumptions made at the commencement of this comparison. However, for the sake of completeness, any such matches have been identified and recorded separately in Fig. 3.9 and Table 3.2. The evaluation statistics described below will be calculated under two conditions. Condition 1 will assume that only the planted sequences constitute 'true' positives whereas condition 2 (denoted below as the 'revised' evaluation) will assume that all sequences that completely match the planted ones constitute 'true' hits. Apart from the planted sequences, 15 sequences that constitute exact matches to the sequences in the training set exist in *random1*. Specificity and precision rate as well as classifier efficiency have been calculated separately whilst taking those sequences into consideration and included for comparative purposes.

Finally, the precision rates obtained from these tests are also arbitrary since they depend on the number of false positive hits, which in turn depends on the size of the scanned sequence library. This is because any and all hits outside the collection of planted sequences will be considered as false. Therefore larger sequence libraries are more likely to contain accidental matches to the planted sequences (or indeed any other high scoring sequence) and all model tested on them will appear to be less precise. This consideration is only relevant under condition 1. The best performing

model, as determined by these evaluations, was used to scan the *D.melanogaster* genome for putative Six4 bss.

It is also worth mentioning that the precision rate was chosen as a criterion over the more commonly used specificity value i.e. the ability of a classifier to distinguish between true and false negatives and minimise the occurrence of the latter [TN / (FP + TN)]. This was done because the performed assessment does not generate negative results in the true sense. Every potential sequence that is not selected is a potential negative result. Because of this, model evaluation will not be presented in the form of a receiver operating characteristic (ROC) curve as is the utilised convention when a classifier system's (in this case model) performance is assessed in response to the alteration of its discrimination threshold (usually a cut-off point) (Metz, 1978; Zweig and Campbell, 1993).

Seeded sequence no.	sequence	Six4 PWM (*align2*) score
1	GTAACCCGAC	11.2
2	GTAACCTGT	8.2
3	-TAACCTGATC*	10.9*
4	GTTACCCGAC	9.4
5	GTCACCTGACAC	9.9
6	GTAACCCGAG	11.2
7	GTTACCCGAA	9.4
8	GCAACCCGAT	9.7
9	GTAACCTGA	10.9
10	GTCACCGGAC	9

* shorter sequences that contain the – symbol rely heavily on the sequence they are embedded in for generating a good score. The assigned score is provisional and required the nucleotide upstream of the embedded sequence to be a G as it was in *random2*.

Table 3.1 Sequences and associated PSSM scores of the 10 planted sequences that were part of the *align2* alignment. Scores are determined by the PWMs generated using both the *align1* and *align2* MSAs.

3.6 PWM generation and evaluation

The following sections will address the operational consideration of constructing and evaluating a PWM based on the SELEX generated data on the binding specificity of Six4.

3.6.1 Generation of a Six4 PWM

Unlike the generation of a HMM described below (section 3.7), the generation of a PWM is an unambiguous process. The only real consideration is the length of the matrix and the number of sequences that will contribute to it. When deciding the length of a TFBS PWM based on an MSA one needs to make an informed decision based on sequence similarity. The inclusion of highly variable positions at the ends of an MSA can dramatically reduce the per position information content of a matrix (see below). With this in mind the motif identified by MEME after the addition of the flanking nucleotides (*align2*) was used as the basis of the PWM. This resulted in a 9 bp long matrix. This size, as well as the base composition of the PWM as expressed as a consensus (GTAACCyGA) was consistent with previously reported SIX4/5 subfamily TFBSs (see chapter 2).

As described previously a PWM is an expression of the likelihood of encountering each nucleotide at each position written as either a nucleotide frequency (a position frequency matrix, PFM, Fig. 3.5) or an integer (Fig. 3.6). A graphical representation of the PFM as obtained through WEBLOGO can be seen in Fig. 3.8.

The pseudocount value for this matrix was set to 0.01 (see section 3.6.3). The role of a PWM as a classifier depends on the weights assigned by it to different positions. The weight matrix can be seen in Fig. 3.7. Pseudocounts are dealt with in more detail in section 3.6.4. As discussed previously the information content of this PWM was determined as being 13.4 bits over a length of 9bp. A statistical evaluation of the TRANSFAC database performed by Fogel et al. (2005) revealed that the average length of the matrices includes in that database 11.9 ± 4.6 nt with a minimum of 2 nt and a maximum of 32 nt. Fogel et al. (2005) do not comment on the average information content of the PWMs in the TRANSFAC database, although my experience suggests that the Six4 PWM compares very favourably to most of the matrices in that database in terms of information content. Information content is dealt with in the following section.

Position	1	2	3	4	5	6	7	8	9	
a	0.0	0.0	0.7	1.0	0.0	0.1	0.0	0.0	0.9	
c	0.0	0.1	0.2	0.0	1.0	0.9	0.4	0.0	0.0	
g	1.0	0.1	0.0	0.0	0.0	0.0	0.1	1.0	0.0	
t	0.0	0.8	0.1	0.0	0.0	0.0	0.5	0.0	0.1	
f.sum	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	9.0
f.max	1.0	0.8	0.7	1.0	1.0	0.9	0.5	1.0	0.9	1.0
f.min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fig. 3.5 Position Frequency Matrix of the Six4 binding sequence

A	0	0	12	17	0	2	0	0	16
C	0	2	3	0	17	15	7	0	0
G	17	1	0	0	0	0	2	17	0
T	0	14	2	0	0	0	8	0	1

Fig. 3.6 Table of the number of instances of each nucleotide observed at each position during SELEX.

Position ;-	1	2	3	4	5	6	7	8	9	
a	-7.4	-7.4	1.0	1.4	-7.4	-0.8	-7.4	-7.4	1.3	
c	-7.4	-0.8	-0.3	-7.4	1.4	1.3	0.5	-7.4	-7.4	
g	1.4	-1.4	-7.4	-7.4	-7.4	-7.4	-0.8	1.4	-7.4	
t	-7.4	1.2	-0.8	-7.4	-7.4	-7.4	0.6	-7.4	-1.4	
w.sum	-20.9	-8.4	-7.5	-20.9	-20.9	-14.4	-7.1	-20.9	-15.0	-136.1
w.max	1.4	1.2	1.0	1.4	1.4	1.3	0.6	1.4	1.3	1.4
w.min	-7.4	-7.4	-7.4	-7.4	-7.4	-7.4	-7.4	-7.4	-7.4	-7.4

Fig. 3.7 Position Weight Matrix of the Six4 binding sequence.

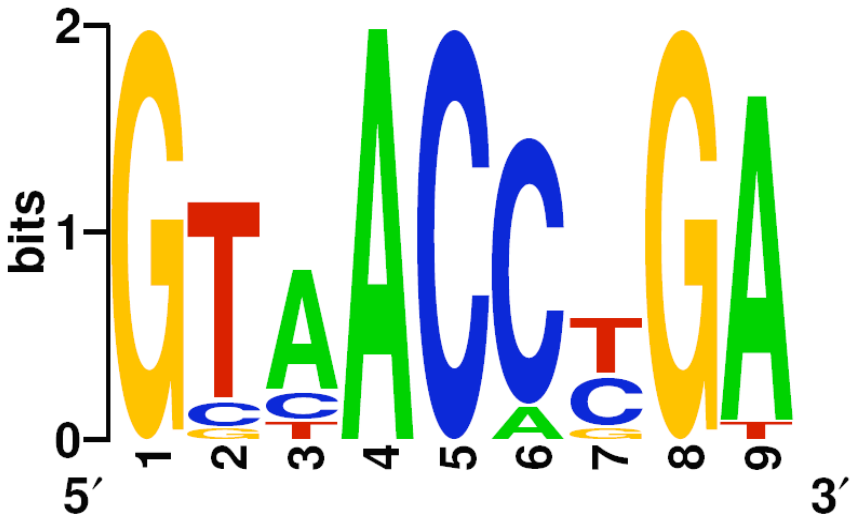


Fig. 3.8 Logo of the Six4 PWM. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each nucleotide at that position. Image generated using WEBLOGO (weblogo.berkeley.edu)

3.6.2 Matrix information content

The information content is essentially a measure of the usefulness of a matrix in identifying a sequence that correctly aligns with the MSA it is based on, rather than generating a false positive (Mount, 2004). It is imperative that a matrix is representative of the sought after target sequences. The quantity of the information inherent in a PWM varies for each column (position) in the motif. This information content is dependant on the background letter frequencies of the studied sequences. The information content is measured in bits. For DNA sequences the information content ranges from 0 to 2 bits with 0 representing an equal probability for all four nucleotides and 2 representing an invariable position. Therefore in the context of a PWM the information content of each position provides a measure of the tolerance for substitutions at each position and the overall information content of a PWM is a measure of its ability to detect genuine binding sequences.

3.6.3 PWM optimisation and Pseudocounts

Algorithms for the optimisation of PWMs have been created. They utilise supplemental data such as ChIP derived data to better inform a PWM (GAPWM, Li et al., 2007). However in the absence of such data, and beyond optimisation based on the stringency values obtained through the use of different threshold parameters, PWMs often have to be manually ‘‘tweaked’’ if one is to achieve the best compromise

between sensitivity and precision rate. This is often done by the addition of pseudocounts to a PWM.

Pseudocounts, as implied by their name, are values that inform the frequency matrix that do not originate from original observations. It is quite conceivable that nucleotides are not represented in the frequency table because not enough sequences were sampled. Pseudocounts are generally used to counter this possibility. Their inclusion in a PSSM is generally accepted to increase its performance (Gribskov and Veretnik, 1996). Additionally they can be used to incorporate more information into PWM that may originate from a different source (in this case the position mutagenesis) and as such can't be normally included in the PWM.

There are various approaches to assigning values to pseudocounts, ranging from the simple option of starting with 1 in each position of a frequency table (Ed Green, MotifBS user's guide 2003) to more informed methods that take the sample size and composition in to account (Henikoff and Henikoff, 1996; Gribskov and Veretnik, 1996). Essentially when one has high confidence in the size and composition of a training set then fewer pseudocounts should be added. Furthermore the addition of a pseudocount value is an operational necessity when generating a PWM. The conversion of a PFM to a PWM and the accompanying conversion of nucleotide frequencies to the weights implies the computation of a logarithm. Therefore values of 0 for non-observed nucleotides must be removed. Given the large size of the training sample as well as its demonstrable homogeneity an informed decision was made to keep the pseudocount value as 0.01 (a value very close to 0). This decision was based on the fact that the effect of the use of different pseudocount values could not be assessed based on the sensitivity and precision rate analysis used to evaluate different methods. This is because the seeded sequences do not incorporate the non observed nucleotide represented by the pseudocounts and as such their addition does not enhance sensitivity but could potentially reduce the precision rate. A more informed decision about the optimal pseudocount value could be made if the identities of true Six4 TFBSs were known and model performance could be objectively measured. As it was the pseudocount value was set so as to not offset the input of genuine counts towards model performance. An additional reason for avoiding the use of pseudocounts based on previous observations was to allow for an unambiguous comparison between the performance of a PWM and a HMM in detecting Six4 TFBSs. The addition of pseudocounts to a HMM would potentially offset a number of

other (unknown) parameters beyond the occupancy of the position for which the pseudocount is reported.

3.6.4 Cut-offs and p-value considerations

As discussed previously the precision rate of a model can be enhanced by setting a cut-off point that essentially eliminated all hits that do not achieve a certain score. As such I have evaluated the performance of the Six4 PWM using a number of classification thresholds (cut-off points). Cut-offs are usually set on the basis of a p-value, or the probability that the background model (based on composition of background sequences) can achieve a score at least as high as the observed one. Given the fact that background models differ, a p-value is not a universal expression of statistic significance. Huang et al. (2004) use a Local Markov Model (LMM) as a means of modelling the properties of background sequences, or the ‘‘null distribution’’ as they call it, as a Markov chain. This algorithm can be used to independently establish a relationship between the p-value generated by PWMs and the sequence analysed, thereby lending a more universal character to the p-value. This approach is essentially equivalent to the use of a null model by HMMER but goes beyond simple sequence composition to take into account local sequence properties (Huang et al., 2004). However, I have been unable to obtain the LMM algorithm from the original authors. In the absence of this, the cut-off points for the Six4 PWM were set based on the weight score achieved by the sequences being scanned. Weight score is independent of background and as such constitutes an unambiguous way controlling a model selectivity and precision rate.

It should be noted that signal-to-noise ratios (precision rates) generated by this comparative analysis are not absolute values and only serve to compare the different sets of variables involved in model design and application. They are products of the specific circumstances under which they were obtained. They do however provide relative information about the performance of different sets of parameters.

3.6.5 PWM evaluation using different cut-off points

The nature of a PWM is such that sequences can only ever achieve certain scores based on their composition. This makes setting weight-based cut-offs an informed decision rather than an arbitrary one. Using the Six4 PWM and assuming a universal pseudocount value of 0.01 these scores or ‘‘states’’ can be the following :8, 8.2, 8.4,

8.5, 8.6, 8.7, 8.9, 9, 9.2, 9.4, 9.6, 9.9, 10, 10.2, 10.9, 11.1 and 11.2 (for reasons of convenience scores below 8 have been omitted, since their use would generate an uninformative number of false positive hits).

Similarly, the scores of the aptamer sequences planted in *random2* are provided in table 3.1. Equipped with this knowledge, one does not really need to use the PWM to scan *random2* to be able to predict the matrix sensitivity under different cut-off points. Rather the cut-off points can be set to obtain a desired sensitivity for this test. The information gleaned from this test is the precision rate (signal to noise ratio) of the matrix using different cut-offs. This allows for a compromise between sensitivity and precision to be made. The cut-off points utilised were 7.2, 8.2, 9.4 and 9.7 corresponding to the scores assigned to planted sequences 10, 2, both 4 and 7 and 8. Higher cut-offs were not used since it was deemed that sensitivity would suffer in response. Numbers of 'true' positive hits achieved under different cut-offs, as well as the total numbers of hits and the matches to the planted sequences present in *random1* are shown in Fig. 3.9. Resulting sensitivities and precision rates calculated using both evaluation methods (see section 3.5) are presented in Figures.3.10 and 3.11.

At this point an informed decision had to be made concerning the optimal cut-off point. This could be achieved in a relatively objective way by choosing the cut-off for which the value of classifier efficiency (sensitivity x precision rate) is the highest. This assessment method can usually provide the best compromise between these two values. Based on this comparison, 9 was deemed to be the most efficient cut-off point out of all the ones that were tested. This was true under both evaluation methods (see section 3.5). Interestingly, the results of the two evaluation methods ranked the other two cut-off points differently in terms of classifier efficiency. However their agreement on the best cut-off point circumvents the need to decide on the most appropriate evaluation method. The following section deals with the construction of multiple hidden markov model based on the same alignment that was used to generate this PWM (*align2*). Sensitivity, precision rate and classifier efficiency are then measured and compared for all the resulting models.

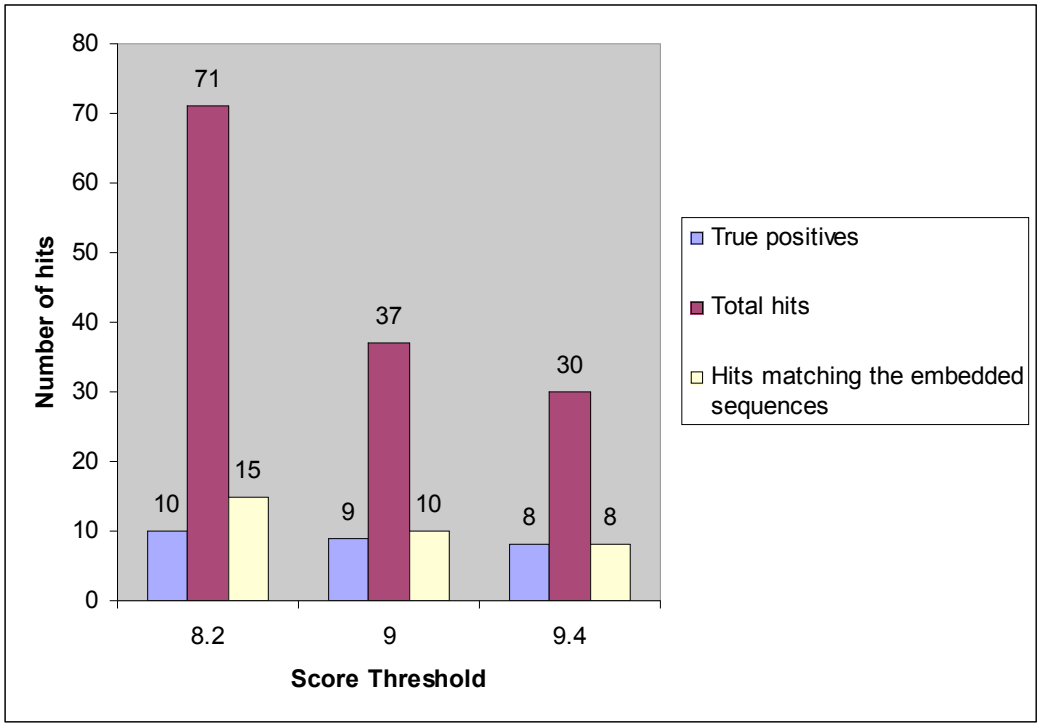


Fig. 3.9 Chart showing the number of true positive hits (those that correspond to the planted sequences), the total number of hits achieved by the Six4 PWM and the number of resulting 'new' hits that match the planted sequences using different score thresholds.

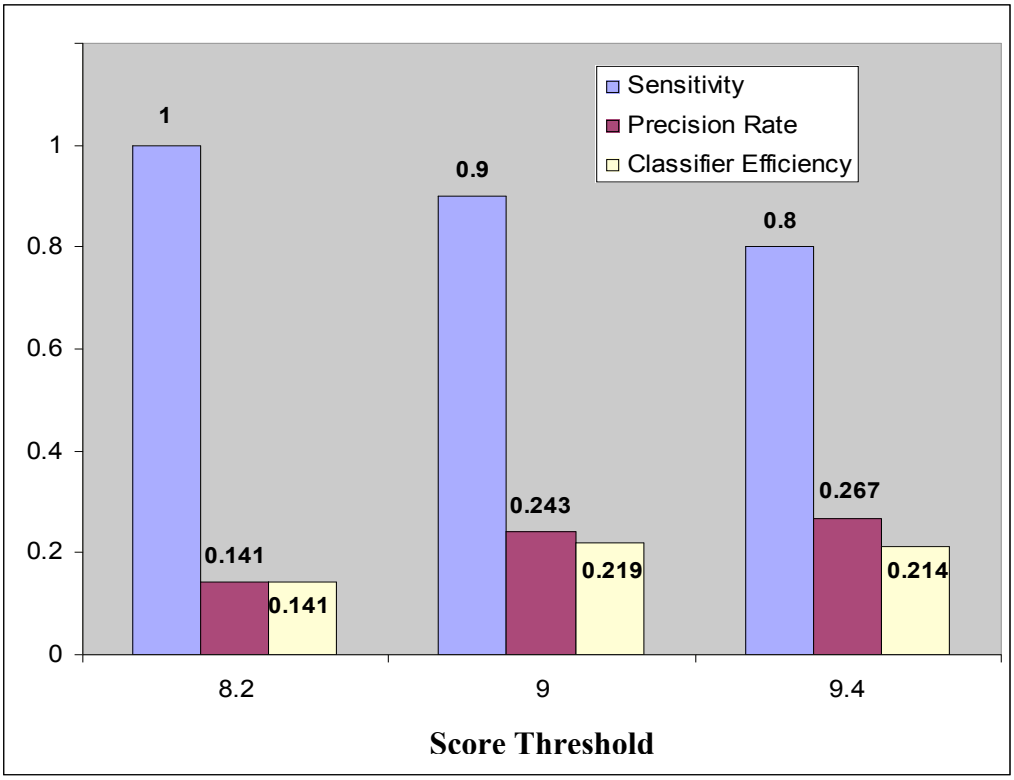


Fig. 3.10 Chart showing the values for sensitivity, precision rate and classifier efficiency as determined by searching the *random2* sequence library using different score cut-off points. These values were calculated using the assumption that matches to the embedded sequences present in *random1* do not constitute 'true' positive hits (condition 1).

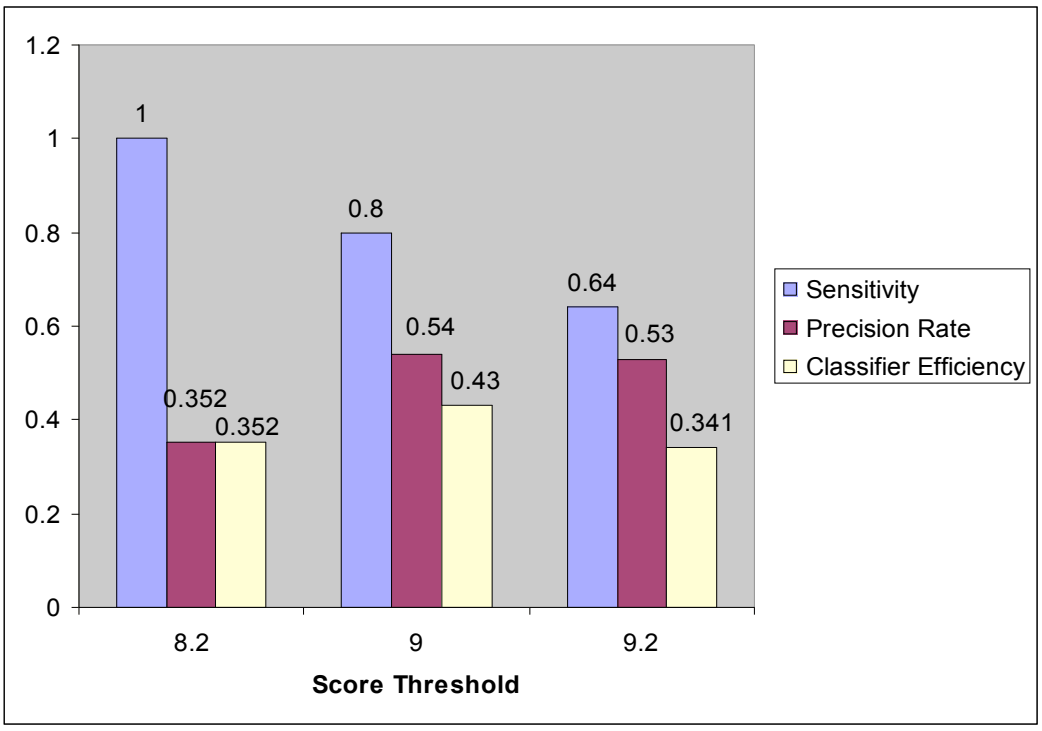


Fig. 3.11 Chart showing the values for sensitivity, precision rate and classifier efficiency as determined by searching the *random2* sequence library using different score cut-off points. These values were calculated using the assumption that matches to the embedded sequences present in *random1* constitute 'true' positive hits (condition 2).

3.7 Hidden Markov Model generation and evaluation

The following subsections deal with the generation and subsequent evaluation of a number of profile hidden Markov models for the detection of putative Six4 TFBSs.

3.7.1 Generation of Hidden Markov Models through HMMER2

The alignment of the 17 consensus conforming isolates (*Align2*)was initially used to build a hidden markov model (HMM) using, both the MAPPER (Marinescu et al. 2005) interface that utilises the HMMER2 utility (Eddy, 1998) as well as HMMER2 directly (to allow for modification of critical parameters). Essentially, the most important difference between the use of the MAPPER interface, where the parameters are fixed, and the default parameters utilised by HMMER2 involves the use of the null model that is used in HMM generation. Essentially MAPPER is just another interface for HMMER2 and was used to gain an insight of the considerations involved in HMM generation through HMMER2. All the models presented in this study were generated

using the HMMbuild utility of the HMMER2.3.2 (latest edition) algorithm (Eddy, 1998).

A number of parameters are used in generating a HMM through HMMbuild. These are discussed below. They essentially constitute the user's input in model generation and help inform the model. They are all included in what is called the null model. A null model is used to calculate log odd scores and is important in generating a HMM. It states the expected background occurrence frequencies of the 4 nucleotide bases and also contains a parameter called p1, which is related to the mean length of a target DNA sequence (mean sequence length or MSL, this is the size of search window used in scanning target sequences). Essentially, if the expected mean length of target sequences is *x*, p1 should be x/(x+1). The default null model for DNA used by HMMER 2.3.2 assumes equiprobability for all 4 bases and a mean sequence length (MSL) of 1000. This MSL is optimised for use with RNA in the default HMMER2 parameters, where sequence lengths can be substantial but is ill suited for use with TFBSs, the lengths of which vary but are generally smaller than 20bp (mean TFBS length in TRANSFAC is 11.9 ± 4.6 nt). Models generated with the MAPPER interface (optimised for TFBS detection) assume a MSL value of 50 (Sean Eddy quoted in Marinescu et al., 2005). Models that assume a smaller MSL value generally generate more hits satisfying an arbitrary cut-off point based on an E value (see below) but are generally less selective (have lower precision rates). It is therefore essential to optimise these values as they have an impact on model performance.

When generating a model under the native HMMER2 environment, the values contained in the null model were substituted for the nucleotide frequency ratios observed in upstream non-coding regions in the *Drosophila* melanogaster Ensembl collection (roughly 0.3 A/T and 0.2 G/C). This is to correct for compositional bias within the model. Additionally , the MSL parameter (and the p1 parameter dependent thereupon) was set to a number of values from 50, as suggested by Sean Eddy in Marinescu et al. (2005), all the way down to 10 (a likely value given the length of both the currently observed Six4 binding sites, as well as the lengths of previously reported Six5 binding sites). This was done in order to establish the ideal value the *x*, and therefore the p1, parameter should take. All the values between 50 and 10 were tested for their ability to detect genuine Six4 binding sites (see below, Table 3.2). The purpose of model testing was to generate a model that generated the highest number of true positive hits and the lowest number of false positive hits. All models and their

respective performances are discussed below. Before discussing model performance the concept of an E value and its utility in HMMs will be discussed.

3.7.2 HMM E value

An HMM search against a given database identifies sequences that match the model and compares them against the scores of sequences used in generating the model. Hits are evaluated on the basis of an E value. The E value represents the expected number of false positives with scores at least as high as the resulting hit. The E value is heavily dependent on the database being searched and is therefore only usable if one knows the database size. Database sizes are provided where necessary. Hits in smaller databases will therefore generate lower E values. To avoid bias, cut-off point should in principle be determined on the basis of hit weight scores (as detailed in section 3.6.4). Accuracy of a HMM can be improved through calibration. Calibration essentially assigns two parameters to a HMM. ‘‘These parameters are the μ (location) and λ (scale) parameters of an extreme value distribution (EVD) that best fits a histogram of scores (based on E value) calculated on randomly generated sequences of about the same length (controlled by a Gaussian distribution) and base composition as the training set’’(Eddy, 1998). Essentially calibration provides the model with an indication of the score a randomly generated sequence is expected to achieve and informs it accordingly. The resulting μ and λ values for the models generated in this study depend on the null model utilised in model generation. Since the background nucleotide frequencies are the same for all experimental models the values of μ and λ depend solely on the value of MSL used in the null model. The relationship between MSL and μ can be seen in Fig. 3.12. Essentially higher μ values indicate a reduced ability of a HMM to discriminate between true and false positives (or rather its tendency to generate an unhelpful number of the latter). All models were calibrated using the same seed (the same value distribution). All resulting models were then tested for their ability to detect the sequences planted in *random2*.

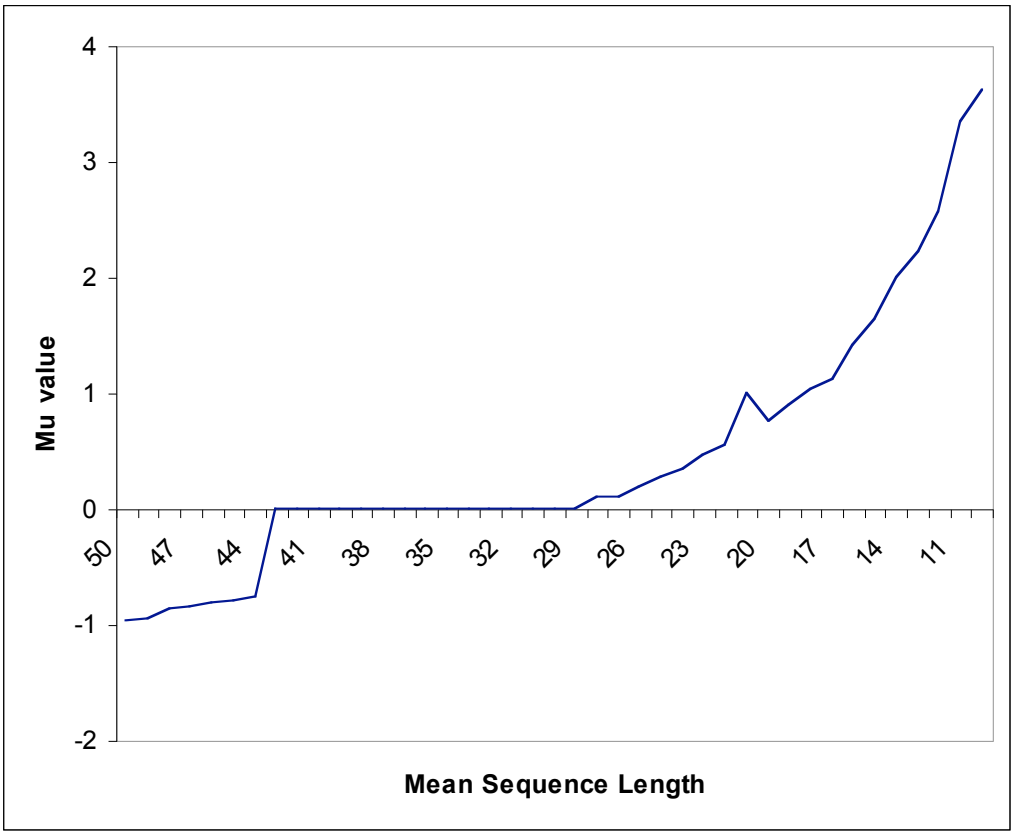


Fig. 3.12 Graph showing the μ value (the location parameter of an extreme value distribution (EVD) that best fits a histogram of scores, based on E value, calculated on randomly generated sequences of about the same length as the MSA the model is built on) in response to the MSL value assumed in the null model of an HMM of the Six4 binding sequence. MSL values range from 10 – 50 bp.

3.7.3 The HMM null model

All models used in this analysis were calibrated using the same seed (essentially the same collection of randomly generated sequences) so as to be directly comparable. μ values for the resulting models can be seen in Fig. 3.12. A high E value (10) was required for detection of true positives when the align1 alignment was used hinting towards a series of weaker models generated with that alignment. Table 3.2 summarizes the number of hits that satisfied the weight cut-off point for false positives (variable, see section 3.7.4) generated by HMM models using different MSL values (50–10) when they were used to query the *random2* sequence library.

An additional consideration was the minimisation of FP hits (often at the expense of TP hits). This consideration is particularly important for whole genome analyses where high FP ratios can easily confound a search.

3.7.4 Model evaluation

Being a scoring model, an HMM classifies hits on the basis of weight (much like a PWM, of which it is a potentially more complicated variant). However obtaining the scoring parameters of a HMM constructed using HMMER2 is not entirely straightforward (the model is after all ‘‘hidden’’). Therefore, all scans of *random2*, using the generated HMMs were ran in duplicate. Initially the scan would be ran using an arbitrary E-value-based cut-off point (E-value=10) for generated hits. The resulting hits were examined and the lowest weight assigned to any of the recovered planted sequences was found. This was then used as the new cut-off point. The number of hits corresponding to the planted sequences (true positives) as well as the total number of calls (true positives + false positives) can be found in table 3.2. In general, the planted sequences scored poorly and most of them were never recovered. All evaluation data (sensitivity and precision rate values) is contained in table 3.2.

Expected (Mean) sequence length	Accurate Calls	All Calls	False positive calls that match the embedded sequences	Precision rate	Sensitivity	Classifier Efficiency
50	1	68	5	0.015	0.1	0.0015
49	1	73	5	0.014	0.1	0.0014
48	1	76	5	0.013	0.1	0.0013
47	1	79	5	0.0127	0.1	0.00127
46	2	81	7	0.0247	0.2	0.00494
45	2	82	7	0.0244	0.2	0.00488
44	2	82	7	0.0244	0.2	0.00488
43	2	83	7	0.0241	0.2	0.00482
42	3	87	8	0.0345	0.3	0.01015
41	2	84	7	0.0238	0.2	0.00476
40	3	90	8	0.0333	0.3	0.01
39	3	95	8	0.0316	0.3	0.00948
38	4	104	9	0.0385	0.4	0.0154
37	4	119	9	0.0336	0.4	0.01344
36	5	124	9	0.0403	0.5	0.0202
35	5	135	9	0.037	0.5	0.0185
34	4	144	9	0.0277	0.4	0.011
33	4	155	9	0.0258	0.4	0.01
32	4	166	9	0.0241	0.4	0.0096
31	5	181	9	0.0276	0.5	0.0138
30	5	200	9	0.025	0.5	0.0125
29	5	223	9	0.0224	0.5	0.0112
28	5	247	9	0.0202	0.5	0.0101
27	5	267	9	0.0187	0.5	0.00935
26	5	279	9	0.0179	0.5	0.00895
25	6	305	9	0.0195	0.6	0.0117
24	3	323	8	0.0093	0.3	0.00279
23	2	348	7	0.0057	0.2	0.00114
22	2	416	7	0.0048	0.2	0.00096
21	1	388	5	0.0026	0.1	0.00026
20	1	442	5	0.0023	0.1	0.00023
19	1	456	5	0.0022	0.1	0.00022
18	1	439	5	0.0028	0.1	0.00028
17	1	458	5	0.0022	0.1	0.00022
16	1	473	5	0.0021	0.1	0.00021
15	0	483	0	*	0	0
14	0	489	0	*	0	0
13	0	494	0	*	0	0
12	0	498	0	*	0	0
11	0	499	0	*	0	0
10	0	500	0	*	0	0

Table 3.2.1 Table of values for precision rate and sensitivity corresponding to the different values for mean sequence length (MSL) used in generating different HMMs. All values were obtained through the use of the *random2* sequence library as detailed in section 3.7.4 and were calculated using the assumption that matches to the embedded sequences present in *random1* do not constitute 'true' positive hits (condition 1).

*These results generated no true positive hits and their Precision rate and classifier efficiency is equal to 0.

Expected (Mean) sequence length	All Calls	Modified accurate calls	Modified Sensitivity	Modified precision rate	Modified Classifier Efficiency
50	68	6	0.24	0.088235	0.021176
49	73	6	0.24	0.082192	0.019726
48	76	6	0.24	0.078947	0.018947
47	79	6	0.24	0.075949	0.018228
46	81	9	0.36	0.111111	0.04
45	82	9	0.36	0.109756	0.039512
44	82	9	0.36	0.109756	0.039512
43	83	9	0.36	0.108434	0.039036
42	87	11	0.44	0.126437	0.055632
41	84	9	0.36	0.107143	0.038571
40	90	11	0.44	0.122222	0.053778
39	95	11	0.44	0.115789	0.050947
38	104	13	0.52	0.125	0.065
37	119	13	0.52	0.109244	0.056807
36	124	14	0.56	0.112903	0.063226
35	135	14	0.56	0.103704	0.058074
34	144	13	0.52	0.090278	0.046944
33	155	13	0.52	0.083871	0.043613
32	166	13	0.52	0.078313	0.040723
31	181	14	0.56	0.077348	0.043315
30	200	14	0.56	0.07	0.0392
29	223	14	0.56	0.06278	0.035157
28	247	14	0.56	0.05668	0.031741
27	267	14	0.56	0.052434	0.029363
26	279	14	0.56	0.050179	0.0281
25	305	15	0.6	0.04918	0.029508
24	323	11	0.44	0.034056	0.014985
23	348	9	0.36	0.025862	0.00931
22	416	9	0.36	0.021635	0.007788
21	388	6	0.24	0.015464	0.003711
20	442	6	0.24	0.013575	0.003258
19	456	6	0.24	0.013158	0.003158
18	439	6	0.24	0.013667	0.00328
17	458	6	0.24	0.0131	0.003144
16	473	6	0.24	0.012685	0.003044
15	483	*	0	0	0
14	489	*	0	0	0
13	494	*	0	0	0
12	498	*	0	0	0
11	499	*	0	0	0
10	500	*	0	0	0

Table 3.2.2 Table of revised values for precision rate and sensitivity corresponding to the different values for mean sequence length (MSL) used in generating different HMMs. All values were obtained through the use of the *random2* sequence library as detailed in section 3.7.4. These values were calculated using the assumption that matches to the embedded sequences present in *random1* constitute 'true' positive hits (condition 2).

*These results generated no true positive hits and their Precision rate and classifier efficiency is equal to 0.

3.8 Classifier comparison discussion

In general, the Six4 PWM using the optimised weight threshold was found to outperform all the generated HMMs in terms of both sensitivity and precision rate, with classifier efficiency rates differing by roughly one degree of magnitude. This observation is true under both evaluation methods. The highest classifier efficiency achieved the PWM and the HMM are 0.219 and 0.0202 respectively for condition 1 and 0.43 and 0.065 for condition 2. Fig. 3.13 provides a graphical representation of the evaluation statistics achieved by the PWM using the optimal cut-off (9) as well as the highest scoring (in terms of classifier efficiency) HMM (MSL=36).

Possible explanations for the difference in efficiency essentially revolve around the fact that the overparameterised nature of a HMM renders it a poor match for the PWMs simplistic efficiency when built on the training set at hand. An HMM will often confuse itself trying to generate parameters based on chance associations that result from the relatively small size of the available training set. These findings do not refute the claims of Marinescu et al. (2005) about HMM superiority over PWMs since this is a comparison based on a single training set. It is however obvious that a PWM is a better tool for detecting putative Six4 TFBSs in a whole genome matrix scan. The following sections will describe such a scan of a number of *Drosophila* sequence libraries using the Six4 PWM.

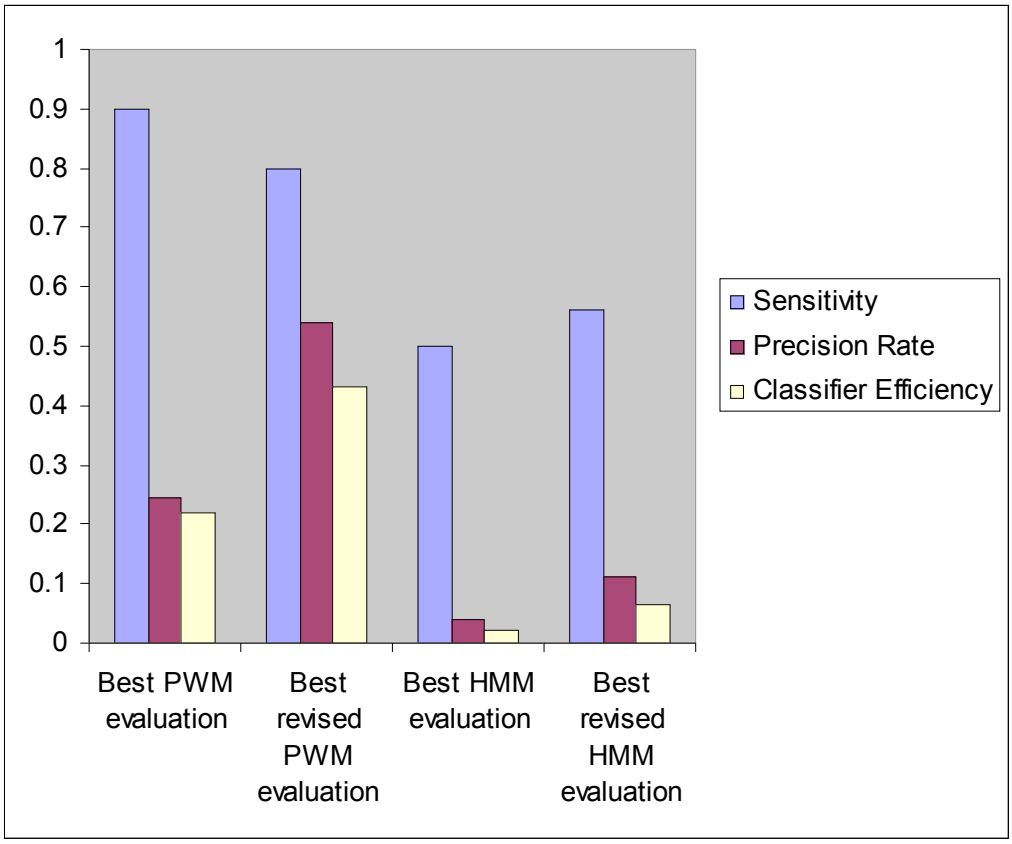


Fig. 3.13 Chart showing the values for sensitivity, precision rate and classifier efficiency achieved by the highest scoring (in terms of classifier efficiency) PWM (cut-off=9) and HMM (MSL=36 or 38 for the revised evaluation, condition 2). These values were calculated using both assumptions outlined in section 3.5. 'Revised' evaluations assume that matches to the embedded sequences present in *random1* constitute 'true' positive hits.

3.9 Whole genome matrix scan

The Six4 PWM was then utilised to probe suspected regulatory sequences in the entirety of the *Drosophila* genome. An initial scan of all the *Drosophila* non-coding sequences (obtained through the UCSC genome browser, http://genome.ucsc.edu/index.html, by removing all Flybase annotated gene sequences from the *Drosophila* genome sequence collection) using the parameters described above revealed 2438 hits to the PWM (data can be made available upon request). A large number of these hits were expected to not correspond to *in vivo* TFBSs. Moreover, since these hits are not associated with any genes (beyond their proximity to annotated loci), their use could potentially bias any resulting gene ontology analyses through the mistaken association of putative TFBSs with unregulated genes. The association of a putative TFBS with a neighbouring gene is not an unambiguous process and potentially opens the door to many false assumptions. The regulatory

sequences that control gene expression are characterised by a degree of uniformity but almost every trend will have its exceptions. It is difficult to definitively assign every putative TFBS to a gene. These considerations, as well as others, are discussed in greater detail in section 3.17.

In light of this, the construction of a putively regulated gene list out of the hits detected in all the non-coding regions was deemed unwise. Such a course of action would exacerbate what is arguably the greatest drawback of the use of PWMs for discovering putative targets of transcriptional regulators i.e. their ability to generate an large number of hits (especially when one considers the number of associations that need to be made between those putative TFBSs and the genes they may regulate). Instead, only a fraction of these hits were associated with their closest neighbouring gene based on the following observations.

Even though TFBSs have been reported in most genomic regions, including coding regions, most of them are known to lie in the vicinity of the transcriptional start (most notably within the 1 kb upstream). This pattern is consistent with their role in the initiation of transcription through direct or indirect polymerase recruitment. Additionally, given the importance of TFBSs in mediating transcriptional regulation, they are often subject to strong evolutionary constraints and are often conserved between closely related species. All this knowledge was used to limit the search space for a matrix scan whilst preserving the potential information content of the scanned sequences. This study has chosen to only make the most likely associations between putative TFBSs and their neighbouring genes and can thus be considered conservative. However, as can be seen by from the results of the gene ontology analyses described in section 3.15, PWM scans (when used in the absence of other data) can confuse analyses based on ontology through high false-positive inclusion rates. It is worth mentioning that the aim of this study is not to identify ‘every’ direct target of Six4 regulation, but to identify some of the most likely candidates. Once some those associations are experimentally validated then the study itself can expanded based on that knowledge.

The Six4 PWM (cut-off set to 9) was used to probe three collections of sequences suspected of containing regulatory elements and therefore being potential hosts of Six4 binding sites. These collections are i) all *Drosophila* gene upstream regions ii) all non-coding genomic regions that show a very high degree of conservation iii) all the experimentally identified *Drosophila* cis-regulatory modules.

Drosophila gene upstream regions consist of all regions up to 1.5kb upstream of an annotated *Drosophila* gene transcriptional start. These regions were truncated when they overlapped with upstream genes. This collection consisted of 19841 entries (circa 21.7 Megabases). The number of entries was based on the number of detected gene transcripts in the Ensembl database (version 48.43b, number of known *melanogaster* genes is 14703) and was obtained using the retrieve sequence utility at http://rsat.scmbb.ulb.ac.be/rsat/ (Thomas-Chollier et al., 2008). Repeats, where present, were masked (through the repeatmasker^{TM} utility, integrated in the UCSC genome browser, A.F.A. Smit, R. Hubley & P. Green, unpublished data, as quoted on http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). Searches were limited to 1.5 kb upstream of annotated genes based on the reasoning that most regulatory information concerning a gene's expression can usually be found within 1 kb upstream of transcriptional start (Arnosti, 2002). Numerous exceptions to this assumption are known to exist but the inclusion of larger sequences would likely generate an uninformative number of false positive hits. This sequence library was designated as *wholegenome* (of course all collections represent sequences culled from the entire genome but this is the largest library used in this study).

Highly conserved genomic fragments have previously been identified for UCSC based on predictions of conserved elements produced by the phastCons program (phastconselement15way multiple sequence alignment, Siepel et al., 2005). Sequences for these fragments were obtained through acquiring a cross-section between all the fragments known to be conserved between the 12 *Drosophila* genomes available through the UCSC genome browser (http://genome.ucsc.edu/index.html, Hinrichs et al., 2006; Kuhn et al., 2007) as well as the *A.gambiae A.mellifera* and *T.castaneum* genomes and the list of all non-coding sequences (based on the fact that coding sequences are usually highly conserved but very rarely contain TFBSs). This collection consisted of 4966 sequences of various lengths and was designated *most conserved*.

Finally the last 2 collections of sequences that were probed consisted of all the publicly available experimentally verified or publicly annotated *Drosophila cis* regulatory elements (CRMs). For reviews of the computational methodology involved in CRM identification see Fickett and Wasserman (2000), Wasserman and Sandelin (2004) and Hannenhalli (2008). These collections were obtained through the REDfly2.0 database (Halfon et al., 2008) that contained 665 sequences (for experimentally verified regulatory sequences, sequence set designated *REDfly*) and

the Open REGulatory ANNOtation database (ORegAnno) collection for *Drosophila* (for the publicly annotated literature derived regulatory elements, sequence set designated *ORegAnno*, Montgomery et al., 2006; Griffith et al., 2008) that contained 2089 sequences. The *ORegAnno* dataset also includes 1,365 sequences reported by Bergman et al. (2005) as being subject to DNAseI footprinting. Both of these libraries were also combined into a single collection designated *CRM.*

Drosophila melanogaster genome annotations were those defined by Flybase (Release 4.3). Simple repeats and repeatmasker regions as defined by the UCSC genome browser were excluded.

The program used to scan these sequence collections was the online interface of patser (Jerry Hertz, web interface designed by Jacques van Helden). The matrix used can be seen in Fig. 3.7. Hits were selected on the basis of a weight cut-off (set to 9) as discussed previously. The pseudocount value was set to 0.01 as described. All generated hits along with associated weights and locations can be found in appendix 3.1

Specifically, the upstream region collection was found to contain 635 hits that satisfied the cut-off (21% less than expected based on the PWM diagnostic run). Since this collection was constructed based on all the available *Drosophila* c-DNAs, some of the upstream regions contained within it correspond to the same locus. These duplicated hits found in over-represented sequences were discarded. This left 365 hits in unique sequences (2.2% of all scanned sequences were hit). Given the large number of genes hit, a complete list of all the hits along with corresponding weights and positions relative to the transcriptional start of the associated gene can be found in Appendix.3.1. Of these sequences, 20 were found to harbour two hits and two contained three hits. The reported molecular functions (as well as the biological processes they were involved in), where available, of these genes did not indicate any obvious link to the Six4 null phenotype. The names of these genes as well as information on their function (where available) are provided in table 3.3.3.

The *CRM* collection generated 44 hits in sequences that are known to regulate 29 different genes (there is some overlap between the reported CRMs and ORegAnno). There was complete overlap between the hits generated from the *CRM* collection (29 hits) and those generated from the *ORegAnno* collection (22 hits). Therefore, all the *ORegAnno* hits have been omitted. The *most conserved* collection generated 2 hits Given the fact that many of the sequences included in this library were shorter than 9bp (and can therefore not include hits to the PWM) it is difficult to determine

whether this library is enriched in hits. These hits as well as their genomic locations and the identities and assigned functions of the genes in their vicinity are described in tables 3.3.1 and 3.3.2.

Gene name	Molecular function	Expression overlaps that of Six4 (as inferred from Flybase expression report and /or BDGP expression data (see section 3.8)	full gene name	Biological Process (where expression overlaps that of Six4 or is unknown)
siz	guanyl-nucleotide exchange factor activity	Yes	*schizo*	CNS development and myoblast fusion
rpr		N/A	*reaper*	Involvement in apoptosis
rho	Peptidase activity	No	*rhomboid*	
kni	Transcription factor	N/A	*knirps*	Numerous processes including muscle development, some Knirps mutants are also known to have gonadal defects (Kirby et al., 2001)
sim	Transcription factor	No	*single-minded*	
Scr	Transcription factor	No	*Sex combs reduced*	
stg	Phospahatase activity	Yes	*string*	Involved in cell cycle
fkh	Transcription factor	No	*fork head*	
ato	Transcription factor	No	*atonal*	
Ubx	Transcription factor	Yes	*Ultrabithorax*	Numerous processes including mesodermal cell fate specification
Ser	Notch signalling	N/A	*Serrate*	

Obp99b	Odorant binding	N/A	*Odorant-binding protein 99b*	
HLHm7	Transcription factor	N/A	*E(spl) region transcript m7*	
slp1	Transcription factor	Yes	*sloppy paired 1*	Numerous processes including segment determination and mesoderm development
salm	Transcription regulator	No	*spalt major*	
pdm2	Transcription factor	Yes	*POU domain protein 2*	NS development
gcm	Transcription factor	Yes	*glial cells missing*	Numerous develomental processes including Glial cell development
dpp	Morphogen activity, transcriptional repressor	Yes	*decapentaplegic*	Numerous developmental processes, regulates Eya
aop	Transcription factor	Yes	*anterior open*	Numerous developmental proceses including muscle development
vg	Transcription regulator activity	N/A	*vestigial*	Wing and other development
Optix	Trascription factor	Yes?	*Optix*	Eye development
gsb-n	Transcription factor	Yes	*gooseberry-neuro*	Segment polarity determination
eve	Transcription factor	Yes	*even skipped*	Various developmental processes
betaTub56D	cytoskeleton component	Yes	*β-Tubulin at 56D*	
Poxn	Transcription factor	N/A	*Pox neuro*	
sc	Transcription activator	? (possible coexpression in the head)	*scute*	

run	Transcription Factor	Yes	*runt*	CNS and eye development amongst others
bi	Transcription Factor	No	*bifid*	
Yp1		No	*Yolk protein 1*	Lipid metabolic activity, Sex differentiation, Vitellogenesis, expressed in the adult fat body

Table 3.3.1 Names and summarised biological function of the genes reported to be under the control of CRMs found to harbour hits to the Six4 PWM. Details about the potential expression overlap between these genes and *Six4* as well as general pertinent gene function information is included (where available, see section 3.8).

Most conserved element ID	Genomic location	Genes in vicinity
dm3_phastConsElements15way_lod=272	**gnl|dmel|3R** 21477104-21477233	*Mir-92b* (unknown function) no expression data
dm3_phastConsElements15way_lod=396	**gnl|dmel|3R** 3300551-3300718	*CG1142* (unknown function) also hit in the *whole genome* scan, no expression data

Table 3.3.2 Genomic locations of conserved (complete conservation between 15 species as observed in the MULTIZ genomic MSA), non-coding, sequences found to harbour hits to the Six4 PWM. Genes in the immediate vicinity (within 2 kb in either direction) along with function and expression data (or lack thereof) are included.

Gene name	Number of hits in upstream region	Molecular function
Adipokinetic hormone-like (Akh)	2	neuropeptide hormone activity
separation anxiety (san)	2	N-acetyltransferase activity
CG14506	2	-
CG14731	2	-
CG15393	2	-
CG17065	2	-
CG2879	2	-
CG31790	2	-
CG32548	2	-
CG33169	2	-
CG33335	2	-
CG5819	2	-
CG6324	2	-
CG6525	2	-
CG6753	2	-
CG6990	2	-
Cyp6t3	2	electron carrier activity; heme binding; iron ion binding
Larval visceral protein D (LvpD)	2	alpha-glucosidase activity; cation binding
CG8877	2	-
CG9203	2	-
CG15373	3	-
crossveinless 2 (cv-2)	3	Molecular function unknown, involved in imaginal disc-derived wing vein specification

Table 3.3.3 Names and summarised molecular function of the genes the upstream regions of which were shown to harbour multiple hits to the Six4 PWM in the *whole genome* scan. Molecular function, where available is included as reported by Flybase.

3.10 Whole genome PWM scan discussion

The scan of the sequence libraries detailed above revealed numerous hits. However, as described earlier, only a small proportion of those are expected to constitute genuine Six4 binding sequences. This reasoning is based on the nature of most scans involving PWMs. It was therefore imperative to discriminate between uninformative and potentially informative hits to allow for an analysis of the results. Specifically, whilst the *CRM* and *most conserved* collections identified 31 genes between them, the *whole genome* library contained 365 unique hits, a large number of which can realistically be expected to be false positives. A GO analysis of all the genes identified by the *whole genome* PWM scan was performed and is included in section 3.15. This analysis revealed no terms with an enrichment score that was higher than 2.5. Additionally, very few terms showed (statistically significant) enrichment. Of those, fewer still were compatible with the role Six4 can be expected to play in *Drosophila* development. This analysis is described, in greater detail, in section 3.15. The inconclusive nature of these findings was attributed to the inclusion of many false positive hits. GO analyses can be fairly sensitive but the the large size of the list of hit-generating genes is likely to reduce their usefulness (especially if the large number of false positive hits that characterise most PWM scans are taken into consideration). This consideration prompted the reduction of the list of potentially regulated genes through filtering of these genes based on their potential association with Six4. This involved the use of a list of genes reported to be differentially regulated in Six4 null background through microarray analysis and therefore potentially subject to Six4 regulation.

3.11 Whole embryo Six4 null microarray screen

A microarray screen of genes showing differential expression in Six4 null embryos ($Six4^{289}$ double mutants, see chapter 1) was performed by Graham Hamilton in 2003 (a then member of Keith Johnson's lab in the University of Glasgow and a collaborator of the Finnegan and Jarman labs). Embryos were collected after 24 hours (and were as such determined to be at various stages of development) and were assayed for differential mRNA expression (Graham Hamilton personal communication). Totals of 525 and 1014 RNAs were shown to be differentially expressed in a positive and negative way respectively (a comprehensive list of these genes has been omitted due to space constraints but is available on request). Due to

the impact of Six4 on *Drosophila* development most of these changes are likely to be secondary effects of the breakdown of the various developmental mechanisms mediated by Six4 and are therefore unlikely to constitute primary regulation targets. For the same reasons the level of differentiation from the wild type expression level of these genes is no indicator of the likelihood of any of these genes being primary targets of Six4. It is worth mentioning that I have limited knowledge of which features were present in the microarray and can therefore not rule out the involvement of Six4 in the regulation of genes not present in the list of reported differentially regulated genes. Moreover, the use of whole embryos may mask the effect of Six4 mediated regulation of proteins in small cell subsets (minute fold-changes may be lost in the noise). For instance, Six4's regulation of expression of a ubiquitous protein in a small number of cells like the somatic gonadal precursors (SGPs) may be missed when the fold-change of that protein's expression is viewed in a whole-embryo context.

The absence of knowledge about which features were present in the microarray as well as the potential unbalancing effect the use of whole embryos may have in reported fold changes, limits the usefulness of the resulting gene lists as a stand-alone source of information. The lists however do contain likely Six4 regulation targets and can therefore be cross-referenced with the list of potential targets generated from a genome-wide matrix scan (*wholegenome* library) to hint towards likely candidates for Six4 regulation.

In light of this these lists were cross-referenced with the list of hits from the *wholegenome* library to yield 32 (3.1%) and 17(3.2%) negatively and positively regulated genes respectively. The names of these genes, as well as data on their potential co-expression with Six4 (where available) can be found in tables 3.4.1 and 3.4.2. The microarray gene lists were found to be enriched in genes containing hits to the Six4 PWM in their upstream region when compared to the rest of the genome (by ~ 40%). This difference is statistically significant (χ^2 statistic is 6.781, $p = 0.01$). These findings are consistent with the expectation for the differentially regulated gene lists to be enriched in Six4 targets. Additionally the fact that both the up-regulated and down-regulated lists are enriched supports the expectation for Six4 to act either repressor or activator depending on the biological context.

Gene ID	Co-expression with Six4 (as inferred from Flybase expression report and /or BDGP expression data	Additional information
CG12310	N/A	N/A
CG3523	N/A	Co-factor binding
CG10864	N/A	Potassium ion transport
CG7714	No	N/A
Rya-r44F	Yes (mesodermal somatic muscles)	Ryanodine receptor, Mutant phenotype consistent with that of Six4 in muscle
Mad	Yes	Transcription factor, role in germ line stem cell division and maintenance, interacts with Eya
stan	No	Membrane receptor
CG15096	N/A	Sodium symporter activity
sage	No	Transcription factor
CG5604	N/A	Ubiquitin protein ligase activity
CG12402	N/A	N/A
ftz	Yes	Numerous processes including gonadal mesoderm development and germ cell migration (Moore et al., 1998_2)
CG9674	N/A	glutamate biosynthetic process
CG9331	N/A	NAD binding
CG7149	Yes	phagocytosis, engulfment
abd-A	Yes	Numerous processes including a major role in gonadogenesis

Cht2	N/A	Chitinase activity
caup	Yes	Transcription factor
T-cp1	Yes	Numerous processes
cdi	N/A	Kinase activity
CG13315	N/A	N/A
gl	No	Interacts with Eya (Tavsanli et al., 2004) maybe a target of Optix
Tsp86D	N/A	N/A
HLHm7	Yes	Transcription factor
Nckx30C	No	compound eye development
B-H1	No	Transcription factor
pk	No	Various developmental processes
olf413	N/A	Copper Ion binding
SCAP	N/A	Protein processing
VhaAC39	Yes	Proton Transport
CG16885	N/A	N/A
Mhc	Yes	Myosin Heavy Chain
CG4306	No	N/A

Table 3.4.1 Table showing the intersection between the genes found to be downregulated in a Six4 null background (D-Six^{289} homozygous) and all the genes found to contain hits to the Six4 PWM in their upstream regions. Details about the potential expression overlap between these genes and Six4 as well as general pertinent gene function information is included (where available).

Gene ID	Co-expression with Six4 (as inferred from Flybase expression report and /or BDGP expression data	Additional information
CG7800	N/A	Protein binding
CG17129	N/A	N/A
CG6685	N/A	N/A
CG8001	No	N/A
CG14317	N/A	N/A
cdep	Yes	actinbinding
Blue	N/A	Pole plasm mRNA localization
CG11819	N/A	N/A
CG12026	N/A	Structural molecule activity
CG4502	N/A	post-translational protein modification
Gdh	Yes	Glutamate dehydrogenase
CG12026	N/A	Structural molecule activity
Unc-119	N/A	N/A
CG15443	N/A	binding
Glued	Yes	Dynactin component homologue (McGrail et al., 1995)
halo	Yes	Microtubule based movement
CG31098	Yes	N/A

Table 3.4.2 Table showing the cross-section between the genes found to be upregulated in a Six4 null background (Six4^{289} homozygous) and all the genes found to contain hits to the Six4 PWM in their upstream regions. Details about the potential expression overlap between these genes and Six4 as well as general pertinent gene function information is included (where available).

3.12 Homologues of identified Six5 targets

Given the high degree of conservation of the transcriptional machinery in all eukaryotes (Aoyagi and Wassarman, 2000; Lee and Young, 2000), the ability of regulatory elements to function in heterologous systems (Kokoza et al., 2001; Mitsialis and Kafatos, 1985; Piano et al., 1999), and the common occurrence of long-distance regulatory elements in metazoans, it is clear that in most aspects *Drosophila* developmental and regulatory pathways are sufficiently related for functional comparisons with other eukaryotes to have some merit. Based on these observations it is reasonable to conclude that regulatory associations known to occur in vertebrates may be maintained in *Drosophila*. In light of this, a detailed knowledge of the regulatory targets of members of the SIX4/5 subfamily in general and of murine Six5 in particular was thought to be important in elucidating downstream targets of its *Drosophila* homologue *Six4*.

Sato et al. (2002) have reported a number of potential interactions that implicate the murine Six5 protein in the regulation of various genes. These authors reported increased expression levels of 23 genes in response to Six5 overexpression. Out of these, 2 genes were previously known to be regulated by Six5 (*Myog* and *ATP1A*). *Myogenin (Myog)* is a gene that encodes a myogenic basic helix–loop–helix (bHLH) protein and is regulated through the MEF3 site, an essential promoter element required for the lineage-specific expression of Myog (Spitz et al., 1998; Ohto et al., 1999). *Igfbp5* is a gene that encodes a component of IGF signalling. The above two genes were shown to be directly regulated by Six5. An additional 21 downstream targets of Six5 are also reported in the same study. Based on the assumption of conserved regulation outlined genes orthologous to those thought to be regulated by Six5 were considered to be likely targets for the Six5 homologue Six4. A search of the Ensembl™ orthologue database and the InParanoid™ utility as well as (http://inparanoid.sbc.su.se/cgi-bin/index.cgi, Remm et al., 2001) revealed 7 orthologous genes in *Drosophila*. All Six5 targets reported by Sato et al. (2002) as well as their *Drosophila* orthologues (where available) are catalogued in table 3.5. Interestingly, none of the upstream regions of these genes (up to position -3000) was found to contain hits to the Six4 PWM. These findings are surprising, since at least in the case of *nau*, the *Drosophila* orthologue of *Myog*, expression and knockdown data as well as information on the general role of *nau* in *Drosophila* development (Paterson et al., 1991; Paterson et al., 1992) is highly indicative of Six4 regulation. A

scan (cut-off lowered to 7.5) of the intronic regions (known to sometimes harbour TFBSs) of the *nau* gene likewise revealed no hits to the Six4 PWM. The same is true of the intronic regions of *CG13830*, the *Drosophila* homologue of the murine *Igfbp5* gene which has been shown to be a target of Six5 (Sato et al., 2002). The locations of the reported TFBSs of murine Six5 near *Igfbp5* with respect to transcriptional start are -70 -2605, -2414 and -1871. No hits were found in the 3kb region upstream of *CG13580* or in any of its introns.

Whilst this observation is curious, this study refrained from widening the search parameters to include larger non-coding regions in the vicinity of these genes to avoid biasing this analysis. This is because, if one looks hard enough then TFBSs (or what looks like them) can be found virtually everywhere.

Finally, a Gene Ontology (GO) analysis of the reported Six5 targets using the DAVID functional annotation clustering utility revealed no enriched GO terms. This analysis was performed in order to gain an appreciation of the GO terms one would expect to be associated with Six4 targets (through functional orthology). No GO terms were found to be significantly enriched in this gene list. The full significance of this is explored in section 3.15 where GO analyses are explained in greater detail.

Murine Gene name	Murine Gene accession number	Flybase ID of reported Inparanoid Orthologue	Reported Orthologue name
MYOG	ENSMUSG00000026459	FBGN0002922	*NAU*
ATP1A1	ENSMUSG00000033161	FBGN0002921	*ATPALPHA*
IGFBP5	ENSMUSG00000026185	FBGN0039054	*CG13830*
IGF2	ENSMUSG00000048583	N/A	N/A
EBF2	ENSMUSG00000022053	FBGN0001319	*KN*
SIX4	ENSMUSG00000034460	FBGN0027364	*SIX4*
SIX2	ENSMUSG00000024134	FBGN0003460	*SO*
DERMO1	N/A	N/A	N/A
MESP2	ENSMUSG00000030543	N/A	N/A
SIM1	ENSMUSG00000019913	N/A	N/A
MDFI	ENSMUSG00000032717	N/A	N/A
FOXD4	ENSMUSG00000051490	N/A	N/A
PURA	ENSMUSG00000043991	FBGN0022361	*PUR-ALPHA*
WNT4	ENSMUSG00000036856	N/A	N/A
SARP1	N/A	N/A	N/A
FZD8	ENSMUSG00000036904	N/A	N/A
PTN	ENSMUSG00000029838	N/A	N/A
GABT3	N/A	N/A	N/A
HTR3A	ENSMUSG00000032269	N/A	N/A
COL9A2	ENSMUSG00000028626	N/A	N/A
KRT1-18	N/A	N/A	N/A
PLTP	ENSMUSG00000017754	N/A	N/A
ST14	ENSMUSG00000031995	N/A	N/A

Table 3.5 Table of the suspected Six5 targets reported by Sato et al. (2002) as well as their *Drosophila* orthologues as obtained through the inparanoid Eukaryotic orthology utility (Remm et al., 2001). Murine gene accession numbers and Flybase IDs, where available, are included. Gene names are included in capitals.

3.13 Analysis of putative target expression

The various sequence libraries scanned with the Six4 PWM yielded a total of 80 genes potentially regulated by Six4 (29 originating from the CRM collection, 2 from the most conserved one, 33 and 17 from the down- and upregulated microarray lists respectively, the *HLHm7* gene featured in both the downregulated and CRM collections and was thus only included once). I have attempted to reduce this number further by comparing the expression patterns of these genes (where available) with that of Six4 and discarding any candidates that showed no overlap. The reasoning behind this was that regulation of genes by Six4 requires their spatiotemporal coincidence. Therefore, when the expression pattern of a candidate target gene can be shown to be incompatible with that of Six4, regulation by Six4 can sometimes be ruled out. Given the context-specific nature of SIX protein mediated regulation one can not rule out the possibility that genes that feature in either the microarray upregulated and downregulated gene lists can constitute direct regulatory targets of Six4. However, since this study is trying to establish direct regulation relationships it is safe to suggest that reported downregulated genes (i.e. those suspected of Six4 activation) will need to have a Six4 compatible expression pattern. One can therefore rule out their role as targets of Six4 because of expression pattern incompatibilities. The same is not true of the upregulated genes since their lack of co-expression might be attributed to Six4 repression. Therefore expression data on all the suspected Six4 targets (where available) was collected and referred to to construct a refined regulation candidate list. The information utilised to this end consisted of the expression data on *Drosophila* genes available on FLYBASE as well as the in situ hybridisation images available through the Berkeley *Drosophila* Genome Project (BDGP) and available imaGO terms (Tomancak et al., 2002). Gene expression was also investigated through the use of the BDGP gene expression query interface. It is often the case that the available data is either incomplete or unconvincing (this is often true of the BDGP in situ images). In those cases the candidate genes were given the benefit of the doubt and were retained in the list of potential Six4 targets. Overall eight genes were excluded from the downregulated list. Where expression of a selected gene is incompatible with Six4 this is clearly indicated in the target gene lists above (tables 3.4.1 and 3.4.2).

This filtering step reduced the number of candidate genes to 72. This gene list is the final product of the Six4 PWM whole genome screen. What follows is a

discussion of some of the considerations made whilst compiling this list as well as some operational and technical limitations. Finally, section 3.15 will make use of this list to perform a gene ontology (GO) analysis of the putative Six4 targets.

3.14 Six4 PWM scan discussion and phylogenetic footprint analysis

The PWM scan detailed above is by no means exhaustive. It is conceivable that in cross-referencing the hit list of the *wholegenome* scan with the list of differentially regulated genes genuine Six4 targets have been overlooked. The Six4 null microarray analysis (and the gene list resulting from it) is also not exhaustive since it can only account for fold-changes in the expression of genes represented on the microarray ChIP (or it may indeed miss certain expression changes due to numerous masking effects). It can be argued that the use of the microarray results may in itself bias the result of the genome scan. This is a calculated risk that one must take when presented with the recurring problem that plagues whole genome matrix scans, namely that they generate a large amount of ‘‘noisy’’ data within which genuine targets are often lost. The same considerations apply to the use of only the regions 1.5 kb upstream of *Drosophila* genes as well as those that are hinted at having biological significance as search targets for a PWM. The expression of many genes is known to be regulated by elements that are very distant to the transcription initiation site (such as the *bithorax* complex, Arnosti, 2003; also the murine Glucocorticoid Response Element, GRE, is known to lie several kb upstream of a start codon, Almon et al., 2005) or may lie in introns (the mesodermal expression of *Six4* itself is controlled through an enhancer in its third intron, see chapter 4, also, expression of the mesodermal factor, Hand, is controlled through an enhancer in the 3rd intron intron of the corresponding gene, Popichenko et al., 2007), 3’ UTRs (37 such elements identified by Stark et al., 2007) or even the coding region itself.

Many of the upstream regions of genes that were hit by the Six4 PWM but are not in the microarray list may contain genuine Six4 binding sites. All this information has been collected for further use (but not directly acted upon). In many ways a complete analysis of the *Drosophila* genome in the strictest sense is beyond the scope of this study (or indeed most other contemporary analyses). The above decisions were made in order to maximise the extracted information without confusing the analysis. As such only the threads that were most likely to yield results were followed. In many ways, the possibilities of a whole genome scan are near infinite and if one is to extract

interpretable data from such an analysis, decisions such as the ones outlined above are generally deemed acceptable.

A phylogenetic footprint analysis of the reported putative Six4 binding sequences was also considered. Such analyses may indicate evolutionary forces acting upon TFBSs as a result of their functional utility and are routinely used for many organisms. Non-coding regions of high interspecific conservation (when compared to the genomic background) are often signs of transcriptional DNA regulatory signals (Siepel et al., 2005). Algorithms that assess the level of this conservation have been developed and range from the fairly simplistic (footprint-discovery at RSAT collects bacterial orthologous genes for a given taxonomical level and discovers conserved elements in their promoters, Janky and van Helden, 2008) to those designed for higher Eukaryotes (Footprinter, Blanchette and Tompa, 2003) and deal with pattern identification and often take transcription factor binding preferences into account (FOOTER, Corcoran et al., 2005). The former are too basic to account for the complexities of most eukaryotic genomes whereas the latter often require knowledge of TFBS behaviour and usually deal with mammalian genomes (FOOTER). Benos et al. (2007) review some of the considerations of phylogenetic footprinting and literature on this topic is visited in greater, though not exhaustive, detail in chapter 4. Interspecific phylogenetic footprinting in conjunction with TFBSs scanning has been successfully used in the past to identify CRMs in various organisms. The reader is advised to refer to Buchanan et al. (2004) for a study in maize, sorghum and rice as well as a very extensive work by Stark et al. (2007) using the alignment of the 12 *Drosophila* genomes for the identification of various evolutionary signatures including transcriptional targets for TFs. The 12 *Drosophila* genome MSA is further used by Kheradpour et al. (2007) to identify targets of 83 TFs.

Discovering footprints in eukaryotic genomes is far from straightforward and it often requires knowledge of the properties of the TFs in question as well as of the genomic region that may house their binding sites and the genes that are suspected of being under their regulation. In the absence of this knowledge, a full phylogenetic analysis of all the genomic regions harbouring Six4 PWM hits in the final list of putative Six4 targets was considered to be infeasible. Such an analysis can not always account for the possibility of a TFBS's relocation within a CRM and does not provide irrefutable evidence of a putative TFBS's utility based in the conservation (or absence thereof) of its sequence. Kheradpour et al. (2007) state that ''many regulatory motifs are too short to guide alignment algorithms and thus may not appear at orthologous

sequences in MSAs''. Individual motifs may relocate through genomic mutations such as insertions or deletions, or through the ''birth of new motifs and the loss of old ones through compensatory mutational changes'' (Kheradpour et al., 2007). Other factors that can account for loss of conservation beyond TFBS relocation include the inability to locate a TFBS in one of the scanned genomes due to improper MSA (an all too common operational risk) or due to interspecific divergence in the function of TFs. Even in the presence of a relatively strong alignment of conserved non-coding DNA that hints towards the presence of a CRM, the identification of individual TFBSs is far from straightforward (Taylor, 2006). Yang et al. (2007) summarise many of the difficulties of phylogenetic footprinting and state that ''such an approach may be too stringent because of the level of degeneracy shown in transcription factor binding site position weight matrices. Due to the degeneracy, there may be only a few bases that need to be conserved across species. Therefore, while a sequence may not show a high level of evolutionary conservation, these sequences may still show high affinity for the same transcription factor''.

In a recent review Hannenhalli (2008) states that ''although reliance on evolutionary conservation is an effective means to reduce the false-positive rate in binding site prediction, conservation is neither a sufficient, nor a necessary condition for biological functionality''. In light of these difficulties, the results of a phylogenetic footprint analysis were not used to discount Six4 regulation candidates from the existing list but can be referred to, to assess whether a putative target is likely to be subject to Six4 regulation.

Conservation of a TFBS between numerous related species provides a dependable indication (not proof) of its functional utility. What will be presented here is a limited phylogenetic footprint analysis of all the hits in the final suspected target gene list. Usually the conservation of individual motif instances over a number of related genomes is assessed based on the Branch Length Score (BLS) i.e. the total branch length of the phylogenetic tree over which the motif is conserved. Such an approach is not used here since no operational decisions are made based on the results of the footprinting analysis. Other methods that combine phylogenetic footprinting and PWM scanning (CONREAL, Berezikov et al., 2004) do exist but are primarily aimed towards vertebrate systems.

Instead, the online interface of EVOprinterHD (a multigenomic comparative tool for the identification of functionally important DNA, Odenwald et al., 2005; Yavatkar et al., 2008; also see Yavatkar et al., 2008 for a description of the identification of 10

known *melanogaster* TFBSs using pair-wise alignments in EVOprinterHD) was used to generate an enhanced 3X-BLAT alignment of the 12 *Drosophila* genomes and assess the conservation of the reported Six4 PWM hits. The repeats that were masked using repeatmasker for the matrix scan described previously were unmasked for this analysis to assist with multiple sequence aligments (EVOprinterHD incorporates a function that prevents repetitive sequences from confusing an alignment). Conservation of putative binding sites between *D.sechellia, D.simulans, D.erecta, D.yakuba, D.ananassae and D.virilis* orthologous DNAs can be seen in table 3.6. This analysis accounts for conservation within the melanogaster group but also includes *D.virilis* (to account for conservation over a long evolutionary distance). Hits are classed as not footprinted when conservation between these species is not complete and as footprinted if conservation is either complete or if changes in individual species would still generate a Six4 PWM hit of weight 9 or higher (permissive mutation).

EVOprinterHD was used to scan all the genomic regions that contain hits to the Six4 PWM that feature in the final suspected target list (70/72 genomic regions, hits in the two entries of the *most conserved* library are footprinted by definition) for phylogenetic footprints that cover the suspected Six4 TFBSs. EVOprinterHD can generate MSAs (called EvoPs) from subsets of species that feature in the UCSC MULTIZ alignment and can therefore sometimes filter out potentially disorientating influences of single divergent sequences. This function allows the experimenter to gain an appreciation of the conservation of a genomic region that might otherwise be masked by the application of extreme (and often unnecessary) search stringency. It is often the case that single divergent or misaligned sequences will confuse an MSA and hide a phylogenetic footprint. At the same time the addition of a phylogenetically distant group (like *D.virilis* in this case) assigns confidence to any reported findings. This last consideration in particular, is important if one takes into account the concept of ‘‘phylogenetic shadowing’’ (term coined by Boffelli et al., 2003). Phylogenetic shadowing refers to the fact that footprinting of distantly related species (like *melanogaster* and *virilis*) is likely to identify only ancient regulatory elements. Shadowing identifies more recently created regulatory elements, through identifying conservation patterns in multiple closely related species (like the members of the melanogaster group). The settings utilised in this analysis account for this possibility by allowing detection of conservation in all but one of the utilised *Drosophila* species. The results of this analysis are summarised in table 3.6.

This analysis is not exhaustive and as such will not be used to inform the gene ontology analysis that is described in the end of this chapter. It does however provide strong indications about the suspected utility of some of the putative TFBSs identified in this screen. Kheradpour et al. (2007) establish a useful methodology for conducting such an analysis in greater detail but that is currently beyond the scope of this study.

Gene Symbol	Phylogenetic Footprinting/Shadowing	Putative binding site sequence conservation
CG3523	Yes	GCAACCTGA
CG12310	No	
HLHm7	Yes	GTAACcGGA
Tsp86D	No	
CG13315	No	
cdi	No	
T-cp1	No	
caup	No	
Cht2	No	
abd-A	No	
CG7149	Yes	GCAACCCGa
ftz	No	
CG12402	No	
CG5604	No	
CG15096	Yes	GTCACCGGA
Mad	Yes	GTTAGCCGA
Rya-r44F	No	
CG10864	moderate	GTTACCCGa
CG7800	Yes	GcAACCTGA
CG17129	N/A	No alignment was possible
CG6685	No	
CG14317	No	
Cdep – 2 hits	Hit 1 moderate Hit 2 moderate	gGAACCCGA gCAACCTgA
Blue	No	
CG11819	No	
CG12026	No	

CG4502	No	
Gdh	No	
CG12026	No	
Unc-119	Yes	GCAACCTGA
CG15443	N/A	
Gl	No	
halo	No	
CG31098	Yes	GTCACCTGA
olf413	Yes	GCAACCTGA
SCAP	No	
VhaAC39	Yes	GTCACCTGA
CG16885	No	
Mhc	Yes	GTAACCGGA
Yp1	No	
bi	No	
run	No	
sc	No	
Poxn	No	
betaTub56D	No	
eve	No	
gsb-n	No	
Mir-309	No	
Optix	Yes	GTAACCCGA *
vg	No	
aop	No	
dpp	No	
gcm	No	
pdm2	No	
salm	No	
slp	Yes	GTCACCTGA
HLHm7	Yes	GTAACCGGA
Obp99b	No	
Ser	Yes	GTAACCTGA

Ubx	**Yes (abx6.8 enhancer)** **No (ventral imaginal disc enhancer)**	**GCAACCTGA**
ato	**No**	
Fkh	**No**	
Stg	**Yes**	**GCaACCTGA**
Scr	**No**	
Sim	**No**	
kni	**No**	
rho	**No**	
rpr	**No**	
siz	**No**	
croc	**No**	

Table 3.6 Phylogenetic footprinting analysis of putative Six4 binding sites. Hits are labelled based on the symbol of the gene next to which they are located. Black capital letters represent bases conserved in all species and coloured bases represent sequences present in all species except *D.sechellia*, *D.simulans*, *D.erecta*, *D.yakuba*, *D.ananassae* or *D.virilis*. Footprinting analysis performed using EVOprinter (see section 3.14).

* Optix shares binding specificity with Six4 (Kawakami et al., 1996; Hu et al., 2008) and is known to regulate itself. It is therefore conceivable that this putative binding site (and possibly some of the others) correspond to Optix binding sites.

3.15 Gene Ontology (GO) analysis

The size of the gene list generated by the genomic scale matrix scan and the expected high incidence of false positive matches necessitated the conversion of this low-level noisy dataset into an informative gene list. To this end, a gene ontology (GO, www.geneontology.org, The Gene Ontology Consortium, 2000-2008) analysis was used to make inferences about the common characteristics of the members of the generated gene list in an attempt to detect patterns consistent with Six4 expression and function.

There are numerous available tools designed to query the GO database. Khatri and Draghici (2005) review 14 of the tools available at the time of publication. Many of the concerns of these authors involve operational differences between these tools. The review does not reach a definitive conclusion but raises a number of concerns about the use of GO analysis tools. Some of the concerns raised by these authors (like the apparent incompatibility of gene IDs) have since been addressed by some of the reviewed programs or others that have been developed since. Most tools are listed on

the GO website (Gene Ontology Consortium, 2008). Currently, DAVID™ (Dennis et al., 2003) is the most cited gene ontology analysis tool. DAVID™ has been used to analyse the list of genes that resulted from the genome-wide Six4 matrix scan. Additionally, Ontologizer2.0™ (http://compbio.charite.de/index.php/ontologizer2.html, Bauer et al., 2008) was used in addition to DAVID™ since it provides an alternative to the more commonly used *term-for-term* analysis in the form of the *parent-child* analysis (Grossman et al., 2007). Bauer et al. (2008) state that most current methods treat each GO term independently and in so doing ignore relationships between GO terms. These authors utilise an approach that takes parent-child relationships into account and thus avoids detecting more specific terms that lie under over-repressented terms. This is possible because ‘‘over-representation of a term is measured with respect to the presence of its parental terms in the set. Our approach comes at no additional computational complexity when compared to the standard approach’’ (http://compbio.charite.de/index.php/ontologizer.over.html). The results of these analyses as well as the utilized parameters are detailed below (sections 3.15.1-3).

The final gene list consisted of 72 genes, 56 of which had GO terms assigned to them (the full list can be seen in Table 3.6 with the addition of the genes *Mir-92b* and *CG1142*). When classified through Ontologizer2.0, genes were screened using a parent-child analysis. By contrast the DAVID™ analysis provides a more classical term-for-term analysis. In both cases the cut-off p-value was set to 0.1. The Bonferroni and Benjamini-Hochberg processes for multiple testing corrections (MTC) were used by Ontologizer2.0™ and DAVID™ respectively. All the terms for the three main GO categories (biological process, molecular function and cellular component) were used.

3.15.1 DAVID analysis

A DAVID™ analysis using the functional classification tool (classification stringency set to medium) revealed one prominent GO cluster in the submitted gene list. The resulting cluster had an enrichment value of 9.68 and consisted of 23/72 genes (effectively 23/56 since 18 genes had no GO terms assigned to them). The genes present in this cluster as well as the GO terms associated with them can be seen in Fig. 3.14. Most of these genes are transcription factors known to mediate development in general and cell differentiation in particular. Moreover some of the genes in this list such as *aop*, *abd-A*, *knirps*, *eve*, *ftz*, *mad* and *dpp* have functions

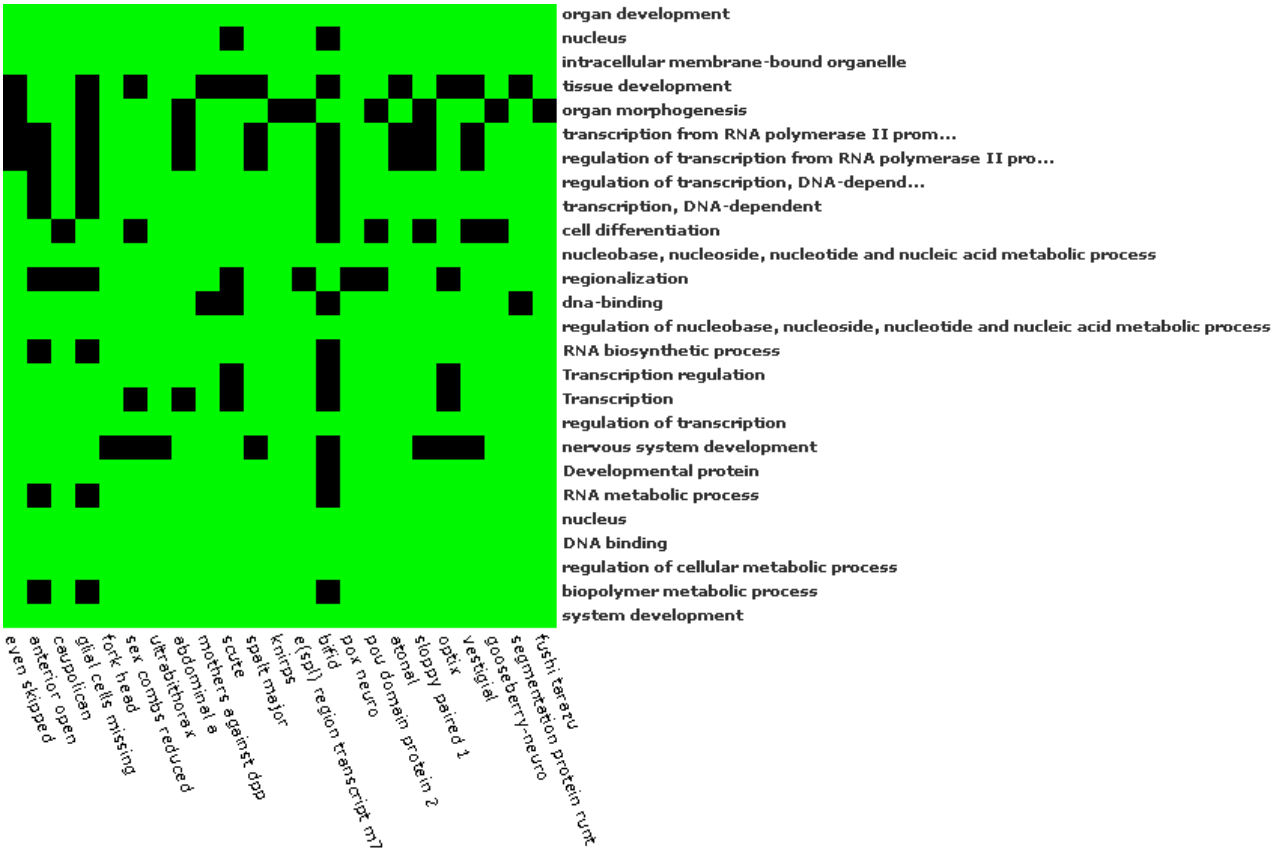
(inferred either directly or through mutant phenotypes) that strongly hint towards association with Six4.

Additionally an analysis of the same gene list through the functional annotation clustering tool revealed numerous gene clusters. Classification stringency was set to ‘‘highest’’ to avoid generating noisy data. Clusters with enrichment scores of 4 or higher as well as their associated p-values are presented herein. Fig. 3.15 shows the most prominent annotation cluster (enrichment score = 13.83). All additional clusters are presented in Appendix 3.2. The significance of the enriched GO categories as well as their potential ramifications for Six4 mediated regulation is discussed in section 3.16.

Finally, a GO analysis of the 365 genes hit in the PWM scan of the *whole genome* collection was also performed. All the over-represented terms with corrected p-values below 0.01 as well as study counts and population percentages are presented in Fig. 3.17. The terms found to be enriched in this analysis showed much lower enrichment scores (highest score was achieved by the term hydrolase and was 2.41) than those achieved by the list compiled through the combined approach (enrichment score of 13.83 for the organ-system-anatomical structure development functional cluster shown in Fig. 3.15). Similarly, corrected p-values were found to be greatly increased when compared to those seen in the refined list (lowest value was 5.70E-05 for hydrolase as opposed to 17E-13 for the cluster mentioned previously). Finally, whilst the highest scoring term (hydrolase) does not show an immediate compatibility with the suspected Six4 function, terms like transcription and NA-binding, as well as developmental protein, were found to be enriched. These terms are also enriched in the refined gene list although their enrichment scores are universally much higher. The associated corrected p-values also echo the trend for higher statistical significance being assigned to the results of the refined list GO analysis. The results of this analysis are useful since they illustrate the ability of the combined approach to potentially successfully reduce the ‘noise’ present in the raw data at a relatively low expense in information content.

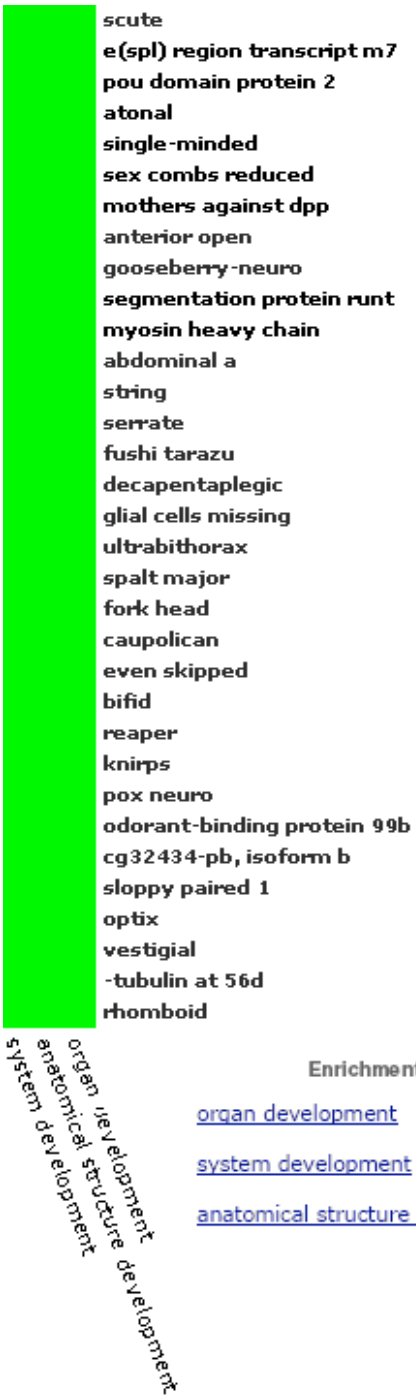
As stated previously the purpose of this study is not to uncover every target of Six4 regulation but to obtain a list of candidates to which high confidence can be assigned. As such the refined list is considered to be a much more useful starting point for any subsequent validation experiments (from a purely statistical point of view) although genes associated with enriched terms in the GO analysis of all the genes hit

in the *whole genome* scan (and not present in the refined list) also constitute interesting candidates for Six4 regulation.



Term	Count	%	p-value	Benjamini
Developmental protein	27	31.4	1.00E-22	5.70E-20
Transcription regulation	22	25.6	2.40E-19	7.00E-17
dna-binding	22	25.6	6.70E-17	2.10E-14
Transcription	20	23.3	1.20E-16	1.60E-14
organ development	34	39.5	1.40E-15	5.00E-12
nucleus	24	27.9	6.80E-14	7.80E-12
system development	34	39.5	5.30E-13	6.20E-10
DNA binding	27	31.4	4.20E-12	2.90E-09
regionalization	20	23.3	2.50E-12	2.20E-09
regulation of cellular metabolic process	29	33.7	4.60E-11	1.80E-08
regulation of transcription	27	31.4	4.20E-11	1.80E-08
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	27	31.4	3.30E-10	7.70E-08
organ morphogenesis	19	22.1	2.00E-09	3.20E-07
tissue development	16	18.6	3.80E-09	5.80E-07
regulation of transcription, DNA-dependent	23	26.7	4.30E-09	6.30E-07
transcription, DNA-dependent	23	26.7	3.00E-08	4.00E-06
cell differentiation	26	30.2	3.70E-08	4.40E-06
nervous system development	20	23.3	1.10E-08	1.60E-06
regulation of transcription from RNA polymerase II promoter	15	17.4	1.90E-07	2.00E-05
transcription from RNA polymerase II promoter	15	17.4	3.10E-06	1.90E-04
RNA metabolic process	25	29.1	2.80E-06	1.90E-04
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	30	34.9	2.10E-05	1.10E-03
biopolymer metabolic process	32	37.2	9.20E-04	2.90E-02
intracellular membrane-bound organelle	35	40.7	1.70E-03	4.50E-01

Fig.3.14 Visual representation of the functional cluster generated by the DAVID™ functional classification tool (used to classify a large gene list into functional related gene groups) using the Six4 putative target gene list. Green squares show where a corresponding gene-term association has been positively reported. The enrichment value for this cluster is 9.68 and it consists of 23 genes. Enrichment ratios, study counts and corrected p-values (using Benjamini-Hochberg MTC) are included in the adjoining table.



Enrichment Score: 13.83		Count	P_Value	Benjamini
organ development	RT	33	5.0E-17	1.7E-13
system development	RT	33	2.0E-14	3.5E-11
anatomical structure development	RT	33	3.3E-12	2.9E-9

Fig. 3.15 Visual representation of the most prominent functional cluster generated by the DAVID™ annotation clustering tool (this tool establishes relationships among the annotation terms present in the gene list) using the Six4 putative target gene list. Green squares show where a corresponding gene-term association has been positively reported. The enrichment value for this cluster is 13.83 and it consists of 33 genes.

The functional annotation clustering results support the findings of the functional classification analysis with terms such as organ development and transcriptional regulation being over-represented. These terms are consistent with the involvement of Six4 in *Drosophila* development. Since terms are clustered, obtaining the highest scoring terms is not necessarily meaningful. Nonetheless, the highest scoring terms (from the highest scoring clusters) as determined by DAVID^{TM} are organ/system/anatomical structure development (consistent with the role of Six4 in such developmental events such as gonadogenesis and muscle founder cell specification and fusion) and transcriptional regulation (once again consistent with both Six4's role in development and its nature as a transcription factor). Additionally, most of the terms that characterise the functional cluster of genes identified by the functional classification tool support the assumption that Six4 will regulate other developmental genes. As described previously, a GO analysis I performed using the 21 reported Six5 target genes (Sato et al., 2002, see section 3.12) using the parameters described in this section has revealed no over-represented terms. As such there are no expectations in terms of what genes Six4 should regulate, although the enriched terms are consistent with the role of Six4 as a mediator of development. The following two sections describe an alternative (and potentially more informative approach) to GO analyses and its implementation with the Six4 putative target list as well as a control analysis performed on the differentially regulated gene lists to rule out initial skewing of this analysis in terms of GO term content.

3.15.2 DAVID^{TM} Analysis of the microarray gene lists

In order to prove that the enrichment of the overrepresented GO terms was a result Six4 regulation and did not reflect a pre-existing trend in the scanned sequences libraries I performed a functional annotation clustering analysis of the two microarray gene lists using DAVID^{TM}. This was done to rule out the possibility of the selected GO terms being over-represented in those two lists and therefore skewing the results of the final GO analysis. Ideally, the same analysis should be performed for the *most conserved* and *CRM* libraries (since they are also contributors to the final putative target list). However this was not done because those two libraries contain references to sequences associated with genes and not the genes themselves. Therefore I did not construct gene lists representing those two libraries to scan for over-represented GO terms.

The gene lists for the up-regulated and down-regulated genes were designated *positive* and *negative* and contained 525 and 1014 genes respectively. The search parameters used for these analyses were those described previously (see section 3.15.1). The GO terms resulting from the putative target gene analysis (organ/system/anatomical structure development and transcriptional regulation) were not found to be enriched in these two lists. However, organ/system/anatomical structure development *child* terms (see section 3.15.3 for a definition of *child* terms) were enriched in these lists. This finding is not surprising given the dramatic effect the absence of Six4 has on numerous developmental processes. Figures 3.16.1 and 3.16.2 summarise these control analyses (see also Appendix 3.2 for less statistically significant gene clusters identified by this analysis). It is therefore reasonable to conclude that the enrichment of the putative target gene list in GO terms consistent with the transcription factor aspect of Six4 function is a product of the regulation of some of those genes by Six4. A more in-depth explanation of the results of these analyses and their possible implications for the developmental processes affected by Six4 (and therefore disrupted in its absence) is possible but it is not the focus of this study.

Enrichment Score: 7.59	Count	P_Value	Benjamini
nuclear replication fork	8	2.6E-8	6.1E-6
replisome	8	2.6E-8	6.1E-6
nuclear replisome	8	2.6E-8	6.1E-6
Enrichment Score: 2.99	**Count**	**P_Value**	**Benjamini**
negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	16	1.8E-4	1.9E-1
negative regulation of cellular metabolic process	16	1.8E-3	4.0E-1
negative regulation of metabolic process	16	3.3E-3	4.2E-1
Enrichment Score: 1.77	**Count**	**P_Value**	**Benjamini**
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	45	6.1E-3	5.2E-1
regulation of cellular metabolic process	45	2.1E-2	7.5E-1
regulation of metabolic process	45	3.8E-2	8.3E-1
Enrichment Score: 1.77	**Count**	**P_Value**	**Benjamini**
chromatin silencing	6	9.0E-3	6.3E-1
negative regulation of gene expression, epigenetic	6	9.0E-3	6.3E-1
heterochromatin formation	6	1.1E-2	6.6E-1
chromatin assembly	6	9.6E-2	9.2E-1
Enrichment Score: 1.58	**Count**	**P_Value**	**Benjamini**
mesoderm formation	5	1.2E-2	6.7E-1
mesoderm morphogenesis	5	1.4E-2	7.0E-1
formation of primary germ layer	5	1.7E-2	7.4E-1
tissue morphogenesis	5	1.7E-1	9.6E-1
Enrichment Score: 1.46	**Count**	**P_Value**	**Benjamini**
ESC/E(Z) complex	3	1.5E-2	3.2E-1
histone methyltransferase complex	3	3.3E-2	5.1E-1
PcG protein complex	3	8.5E-2	7.6E-1
Enrichment Score: 1.18	**Count**	**P_Value**	**Benjamini**
pole plasm oskar mRNA localization	5	4.0E-2	8.3E-1
pole plasm mRNA localization	5	5.3E-2	8.6E-1
pole plasm RNA localization	5	5.7E-2	8.8E-1
pole plasm assembly	5	8.3E-2	9.1E-1
intracellular mRNA localization	5	1.2E-1	9.3E-1

Fig. 3.16.1 List of the most prominent functional clusters generated by the DAVID™ annotation clustering tool using the Six4 null upregulated gene list (positive). Enrichment scores for the selected GO terms are provided. ‘‘Count’’ denotes the number of genes annotated to the corresponding GO term in the study set. P-values prior- and post Benjamini-Hochberg MTC adjustment are provided. Only the first cluster is statistically significant though the rest are included for the sake of completeness.

Enrichment Score: 8.39	Count	P_Value	Benjamini
cell projection organization and biogenesis	55	4.0E-9	5.6E-7
cell projection morphogenesis	55	4.0E-9	5.6E-7
cell part morphogenesis	55	4.0E-9	5.6E-7
Enrichment Score: 5.43	**Count**	**P_Value**	**Benjamini**
neurite morphogenesis	42	3.2E-6	2.1E-4
neurite development	42	3.2E-6	2.1E-4
neuron morphogenesis during differentiation	42	3.2E-6	2.1E-4
neuron development	42	5.7E-6	3.4E-4
Enrichment Score: 4.65	**Count**	**P_Value**	**Benjamini**
apical junction assembly	12	1.2E-5	6.1E-4
intercellular junction assembly	12	2.3E-5	1.0E-3
intercellular junction assembly and maintenance	12	4.2E-5	1.8E-3
Enrichment Score: 3.64	**Count**	**P_Value**	**Benjamini**
leg morphogenesis	12	2.0E-4	7.6E-3
limb morphogenesis	12	2.5E-4	9.3E-3
limb development	12	2.5E-4	9.3E-3
Enrichment Score: 3.43	**Count**	**P_Value**	**Benjamini**
epidermal cell differentiation	10	3.2E-4	1.1E-2
non-sensory hair organization and biogenesis	10	3.2E-4	1.1E-2
hair cell differentiation	10	3.2E-4	1.1E-2
epidermis morphogenesis	10	5.4E-4	1.7E-2
Enrichment Score: 3.27	**Count**	**P_Value**	**Benjamini**
ATPase activity, coupled to movement of substances	25	5.0E-4	8.4E-2
ATPase activity, coupled to transmembrane movement of substances	25	5.0E-4	8.4E-2
hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances	25	6.0E-4	8.6E-2
Enrichment Score: 3.13	**Count**	**P_Value**	**Benjamini**
myoblast development	10	6.9E-4	2.0E-2
myoblast maturation	10	6.9E-4	2.0E-2
myoblast differentiation	10	8.7E-4	2.4E-2
Enrichment Score: 2.47	**Count**	**P_Value**	**Benjamini**
multicellular organismal aging	13	3.4E-3	6.6E-2
determination of adult life span	13	3.4E-3	6.6E-2
aging	13	3.4E-3	6.6E-2

Fig. 3.16.2 List of the most prominent functional clusters generated by the DAVID™ annotation clustering tool using the Six4 null downregulated gene list (negative). Enrichment scores for the selected GO terms are provided. ‘‘Count’’ denotes the number of genes annotated to the corresponding GO term in the study set. P-values prior- and post Benjamini-Hochberg MTC adjustment are provided.

Term	Count	%	p-value	Benjamini
hydrolase	36	10.1	1.00E-07	5.70E-05
Developmental protein	20	5.6	1.90E-06	5.40E-04
nucleus	24	6.8	3.00E-05	5.60E-03
dna-binding	17	4.8	4.30E-05	6.10E-03
Transcription regulation	14	3.9	1.20E-04	1.40E-02
atp-binding	18	5.1	4.00E-04	3.70E-02
Transcription	13	3.7	4.70E-04	3.80E-02
nucleotide-binding	20	5.6	8.10E-04	5.60E-02
receptor	12	3.4	1.50E-03	9.20E-02

Fig. 3.17 List of the statistically significant (corrected p-value below 0.01) annotation terms identified by DAVID™ using the list of the 365 genes hit in the PWM scan of the *whole genome* collection. Enrichment scores for the selected GO terms are provided. ‘‘Count’’ denotes the number of genes annotated to the corresponding GO term in the study set. The percenentage of the study count that the annotated terms represent is also included. P-values prior- and post Benjamini-Hochberg MTC adjustment are provided.

3.15.3 Ontologizer2.0™ analysis of the Six4 putative target list

Ontologizer2.0™ (Bauer et al., 2008) utilises a different approach to DAVID™ when classifying GO terms that are over-represented in analysed study sets. It makes use of the *parent-child* approach which ‘‘reduces the dependencies between the individual term’s measurements, and thereby avoids producing false-positive results owing to the inheritance problem’’ (Grossmann et al., 2007). The ‘‘inheritance problem’’ is based on the concept of *parent* and *child* GO terms. *Parent* terms are generic overarching terms that then give rise to more specific *child* terms (in the way the GO term primary metabolism is the *child* of metabolism which in turn is the *child* of physiological process). In a *term-for-term* approach when a gene is annotated to a term it is also annotated to the less specific *parents* of that term. These associations, however, are not taken into account and can often lead to the identification of false

positive results. The *parent-child* approach informs an analysis through these interdependencies and avoids such pitfalls (Grossmann et al., 2007, have shown *parent-child* procedures to outperform *term-for-term* approaches using MTC as well as real data sets).

Ontologizer2.0™ was used on the Six4 study set using a parent-child intersection analysis method (the more stringent of the two *parent-child* approaches). The results are summarised in Table 3.7 and Fig. 3.18. It is worth mentioning that Ontologizer2.0™ also allows for the more traditional *term-for-term* analysis. When using this parameter the results resembled those generated by DAVID™ extremely closely (and have therefore been omitted).

The highest ranking term was found to be transcriptional regulator activity (Adj-p-Value is 8.55e-14). The genes in the study set that are associated with this term are *HLHm7, Mad, optix, poxn, Scr, Ubx, abd-A, aop, ato, bi, caup, dpp, eve, fkh, ftz, gcm, gsb-n, kni, pdm2, run, salm, sc, sim, slp1* and *vg*. Other terms consistent with transcription factors and their role in development were also over-represented (such as gene expression and DNA binding). Finally, terms consistent with development such as biological regulation, developmental process and cell fate commitment were also enriched.

Go term name	P-Value	Adj.P-Value	Pop.Count	Study Count
transcription regulator activity	1.04e-16	8.55e-14	690	25
biological regulation	2.79e-12	2.29e-09	1967	33
developmental process	1.12e-11	9.22e-09	2357	35
multicellular organismal process	4.15e-11	3.40e-08	2619	36
metabolic process	2.86e-07	0.000235	4972	43
DNA binding	6.49e-07	0.000533	939	22
cell fate commitment	3.21e-06	0.00263	235	15
oenocyte differentiation	4.15e-06	0.00341	7	4
organelle	1.47e-05	0.0121	3223	33
gene expression	3.46e-05	0.0284	1722	25
binding	4.67e-05	0.0384	5745	42
cell	8.19e-05	0.0673	5994	45
pattern specification process	8.38e-05	0.0688	451	18
reproduction	8.60e-05	0.0706	876	14
sequence-specific DNA binding	8.77e-05	0.0720	229	14
cellular process	9.63e-05	0.0790	6658	46
RNA biosynthetic process	9.95e-05	0.0816	714	22

Table 3.7 Results of Ontologizer2.0™ parent-child intersection analysis. All terms with assigned p-values lower than 0.1 are shown. Yellow fields represent molecular function whereas pink and green fields represent cell compartment and Biological process terms respectively. Field intensity is proportional to term ranking. P-values prior- and post Bonferroni MTC adjustment are provided. Population and study counts state the number of genes annotated to the selected GO term in the whole genome and the study set respectively. *Parent-Child* relationships between the terms on this table are presented in Fig. 3.18.

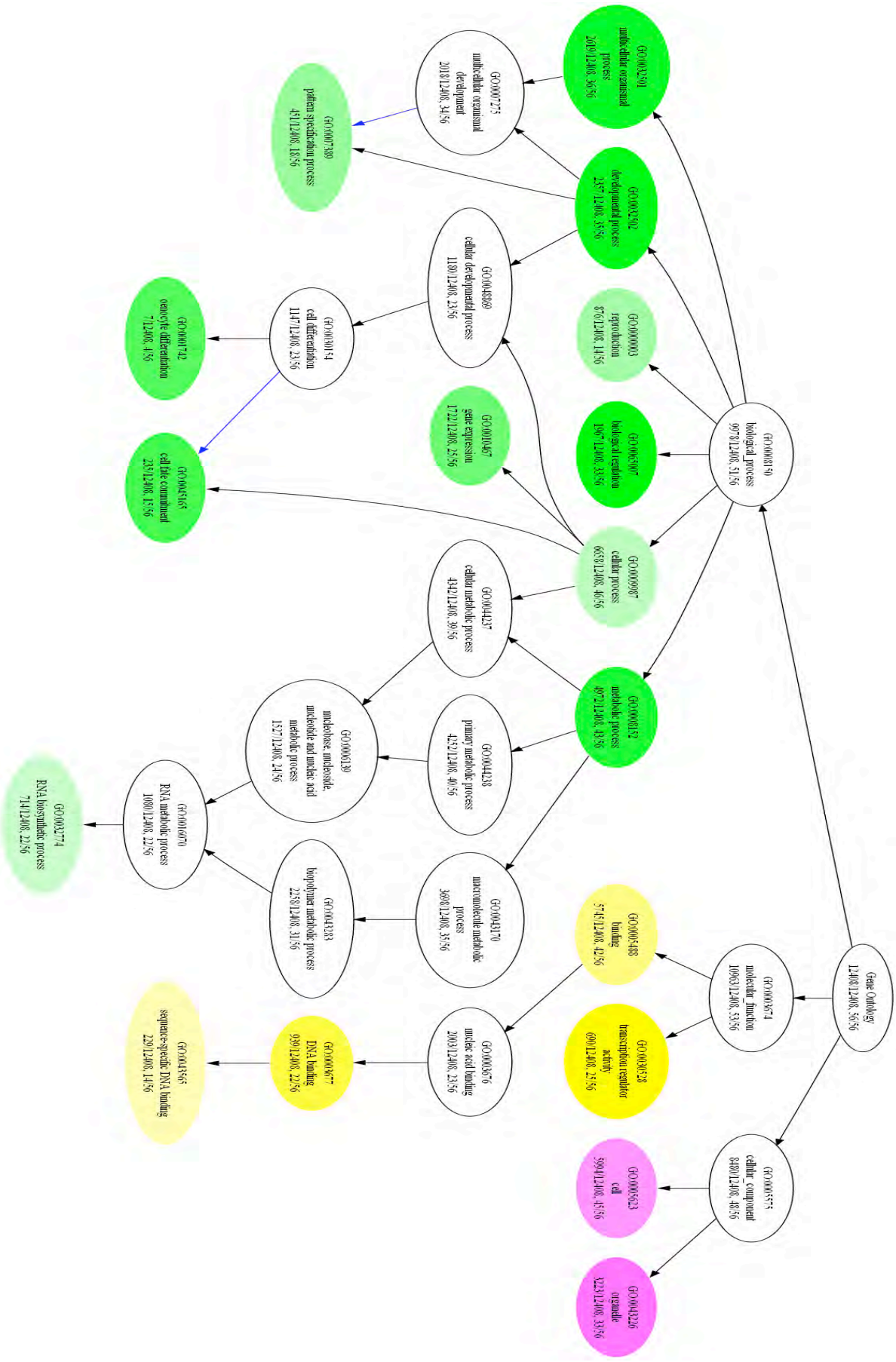


Fig. 3.18 Graph presenting the relationships of the enriched terms detected in the Six4 study set. Arrows are used to connect *parent* and *child* terms. Colour coding as in Table 3.7.

3.16 GO analysis discussion

There is partial overlap between the lists of enriched terms generated by the DAVID™ and Ontologizer2.0™ tools. Existing evidence would assign greater confidence to the results generated by Ontologizer2.0™ due to the more discerning approach it utilises. The MTC approaches utilised by the two analyses also differ. Ontologizer2.0™ makes use of the Bonferonni approach to adjust p-values and correct for the occurrence of false positives. The Bonferonni approach is extremely conservative and can thus often report false negatives and therefore dismiss potentially relevant observations. In contrast the Benjamini-Hochberg approach utilised by DAVID™ is less stringent and thus allows for the potential inclusion of false positives (Camargo et al., 2008). In light of this the results produced by Ontologizer2.0™ will be considered more dependable although the exclusion of a term from the Ontologizer2.0™ results could potentially be attributed to its overly stringent MTC approach. Camargo et al. (2008) state that ‘‘it (the Bonferonni method for MTC) can lead to TypeII (i.e. false negative) errors of unacceptable levels’’.

However, both analyses seem to be in relative agreement over the nature of the enriched GO terms in the study set. Both analyses suggest that genes involved in regulation of transcription and development feature prominently in the Six4 study set. No terms that were associated with the process of transcriptional regulation were found to be over-represented in the differentially regulated microarray gene lists. Given the role of Six4 in development (as evidenced by both its mutant phenotype and the existing knowledge on its orthologues) the regulation of other transcription factors by Six4 is highly likely. This finding hints towards some the genes annotated for this term being subject to Six4 regulation. These results confirm the putative utility of the Six4 candidate target list and suggest subsets of particularly promising regulation candidates (those associated with regulation of transcription, listed in section 3.15.3). These results as well as all those generated from the refined genome-wide Six4 matrix scan are discussed in the following section.

3.17 Discussion and concluding remarks

This chapter addresses the issue of utilising the SELEX generated data to the greatest effect in order to generate a discerning instrument for detecting putative Six4 binding sites. Hidden Markov Models and Positional Weight Matrices were considered for this purpose since they are the most widely used methods for

constructing sequence modelling-based classifiers. The use of Bayesian networks provides an alternative approach to this but was not considered due to the lack of algorithm availability. I have tested both HMMs and PWMs for their ability to detect putative Six4 binding sequences.

Both model types (HMMs and PWMs) are expressions of a multiple sequence alignment that captures the information contained within the SELEX results. As such MSA generation is an important step in model building. I have discussed different approaches to MSA generation and have concluded that the nature of the SELEX process necessitates a step of motif elicitation from the SELEX aptamer library if a useful MSA is to be constructed. A 9bp long motif common to 17 members of the aptamer library was detected and used as the core of the MSA that would inform all generated models. This motif was present in ~40% of the isolated aptamers. This fact was attributed to the inability of SELEX to completely saturate an aptamer pool with ideal Six4 binders. The possibility of Six4 having two divergent binding sequences was considered. However, the lack of homogeneity in the remaining 60% of the isolate pool members and the fact that no such claim had previously been made for any SIX protein family members reduce the likelihood of this possibility. Moreover, the isolated motif showed great similarity to the previously reported Six5 binding sequence. Given the high sequence conservation between the Six5 and Six4 DNA binding specificity-conferring homeodomains this was expected. I am therefore confident that the MSA based around this motif captures Six4 DNA- binding properties and was therefore used to inform a number of binding site detection models.

Both HMMs and a PWM were compared based on their ability to detect putative Six4 binding sites and the ability to reduce the number of reported false positive matches. The resulting evaluation reached a number of conclusions summarised below. These conclusions are based on the use of these models with the Six4 SELEX MSA and as such do not necessarily apply to PWMs and HMMs in general. HMMs have previously been considered as being superior to PWMs (Marinescu et al., 2005) because of their ability to model the higher-order properties of a binding site (a feature lacking in PWMs). A drawback common to both HMM- and PWM-based approaches is that they are only as good as the training data they are based on. Inaccuracies and/or bias within the training data or in the alignment will be captured in the resulting model. PWMs are better at compensating for that than HMMs where

models tend to be over-fitted with (often unnecessary) parameters. An evaluation of 50 different HMMs and a number of weight-based cut-off points for a Six4 PWM have shown the PWM to outperform all HMMs in terms of sensitivity and precision rate. Based on these findings the Six4 PWM was determined as being the best method for detecting putative Six4 binding sites in the *Drosophila* genome. The resulting PWM compares favourably, in terms of information content, to most of the PWMs available through the TRANSFAC™ and the JASPAR™ (Bryne et al., 2008) databases.

This PWM was utilised to scan the upstream regions (1.5kb) of all *Drosophila* genes as well as all *Drosophila* CRMs annotated in the REDFly2.0 database and the non-coding regions shown to be highly conserved between the 12 *Drosophila* species (MULTIZ genome wide alignment) for putative Six4 binding sites. This reduction in search space was an informed decision that was made to avoid generating a large amount of false positive hits. It can be argued that this selective search excludes genomic regions that may harbour genuine TFBSs. I have decided that this search space reduction is acceptable given the existing knowledge on CRMs and their likely locations (see reviews by Wray et al., 2003; Levine and Davidson, 2005; Istrail and Davidson, 2005 and Arnosti, 2003 for more insight on this). This scan revealed hits in 365 different genes. Based on previous experience of PWM scans a large number of these hits were expected to be false positives. In light of this, the list of hits in *Drosophila* upstream regions was cross-referenced to a list of genes found to be differentially regulated in Six4 null mutants in a microarray hybridisation analysis. This was deemed acceptable since that list was considered to be enriched in Six4 direct regulation targets. 32 and 17 hit-harbouring genes were found to be down- and upregulated respectively. All additional hits were catalogued for future reference but their regulation by Six4 was not investigated. Additionally, the reported wild type expression patterns of all genes found to be down-regulated in the microarray analysis of Six4 null mutants (where available) were compared to that of Six4. These genes were candidates for Six4 up-regulation and as such the lack of an expression pattern overlap would dismiss the possibility of an interaction. 8 genes were found to have incompatible expression patterns. This reduction in search space reduced the number of putative Six4 regulated genes to 72.

The hits in the potential regulatory regions of these genes were tested for phylogenetic footprinting between 6 *Drosophila* genomes using the EVOprinter™

footprinting detection utility. A number of footprints were detected and the identities of these as well as the extent of interspecific conservation of the putative Six4 binding sites is summarised in table 3.6. However, detection of phylogenetic footprinting is not straightforward and the absence of binding site conservation does not lessen the potential significance of a hit. There are numerous variables including the potential rebirth of a binding site elsewhere, phylogenetic divergence or even inadequate knowledge of a transcription factor's DNA-binding specificity that can account for the loss of binding site conservation between species. As such I have refrained from making any interpretations based on the potential phylogenetic footprinting (or lack thereof) of putative binding sites. In spite of this fact, binding site conservation may serve as a strong indication of transcriptional regulation when testing potential regulatory associations. These findings have not been used to limit the list of Six4 target candidates any further but are referred to later when suggesting potential experiments that can be performed in the wake of this study.

Finally, the list of the 72 Six4 candidate genes as well as all 365 genes hit in the whole *genome sequence* library scan were used as study sets in a gene ontology (GO) analysis. This analysis was performed using the *parent-child* procedure (in the case of the refined list) as well the more traditional *term-for-term* method. The *parent-child* approach is relatively recent addition to the GO analysis methodology. Grossmann et al. (2007) provide credible evidence supporting the claim that this process may indeed outperform *term-for-term* analyses (such as the one typically performed by DAVID^{TM}). Both analyses showed significant enrichment of the refined putative Six4 target list in genes involved in development and transcriptional regulation. The same is true (albeit in a less pronounced fashion) of the *whole genome* gene list. However, given the higher statistical significance of the refined gene list analysis results, that list would provide a much more suitable starting point in the search for direct Six4 regulation targets.

The term ''transcriptional regulator activity'' (GO:0030528) in particular was shown to not be over-represented in the gene list generated by the Six4 microarray analysis but was over-repressented in the target set. The same is true of ''developmental process'' (GO:0032502) although several of it's *child* terms were found in the microarray list. Both of these terms are consistent with the role of Six4 as a transcription factor with an involvement in such developmental processes as gonadogenesis and skeletal myoblast fusion. A similar analysis using the list of

reported murine Six5 targets revealed no over-represented terms. As such there is no basis for evaluating the GO results beyond inferring potential roles for Six4 targets. The following section suggests a number of experiments that can be performed to investigate the findings of this study. Additionally a list of the most promising candidates for Six4 regulation is included based on the findings presented herein as well as additional knowledge on the function of potentially regulated genes. In many ways the approach utilised in this study is reminiscent of that used by Ostrin et al. (2006) to conduct a genome-wide search for novel direct targets of Eyeless (Ey) in that it is a combinatorial approach utilising *in vitro*, *in vivo* and *in silico* techniques.

3.18 Potential future experiments

As mentioned previously, *in silico* target detection approaches are no substitute for *in vivo* observations. The results of this study are not truly independent but can be of great assistance in informing *in vivo* validation experiments. The data presented herein is by no means exhaustive since this study has aimed to reduce experimental noise by ‘‘mining’’ the genomic regions that were most likely to harbour TFBSs. As described previously, binding sites have been reported in genomic regions not touched by this study. The search space can be expanded to include other non-coding (or even coding regions). However doing so could widen the search space to such an extent that meaningful data might be lost in the noise.

The range of methods used in identifying potential Six4 targets (matrix scan, microarray analysis, footprinting, expression and GO analysis) assign a high degree of confidence to these results. The enrichment in putative Six4 targets of genes known to be differentially regulated in Six4 mutant background strongly suggests that at least some of the targets identified by that analysis are directly regulated by Six4. The results of the final GO analysis in particular identify several transcriptional regulators that could be under Six4 regulation.

Experimental validation can be conducted in a number of ways. Provided antibodies for the proteins in question are available, antibody staining of Six4 null embryos (as well as wild type ones) may provide preliminary results on potential regulation. Given the presence of microarray screen results however the outcome of some of these experiments can be anticipated.

Additional experiments involving manipulation of suspected enhancers can be used to corroborate or disprove the proposed relationship of putative targets to Six4. The driving of reporter constructs in particular by (relatively short) genomic sequences suspected of harbouring Six4 binding sites and the subsequenct removal of the binding sites within those sequences (promoter bashing) can provide more definitive information on the function of a suspected enhancer.

This study lays the groundwork for further analyses and and brings us closer to elucidating the developmental role of Six4. However, I will purposefully refrain from speculating on the potential regulatory function Six4 could perform in light of these findings in the absence of experimental validation.

APPENDIX 3.1. List of hits in the *whole genome* sequence library:

Gene name	Matrix Score	Start	End
CG10031-PA|CG10031-RA	9.40	1483	1491
CG10052-PA|Rx-RA	10.94	358	366
CG10102-PA|CG10102-RA	9.40	253	261
CG10129-PA|ndl-RA	10.94	1347	1355
CG10148-PA|CG10148-RA	9.67	296	304
CG10212-PA|SMC2-RA	10.04	1141	1149
CG10215-PA|Ercc1-RA	10.04	1442	1450
CG10233-PA|CG10233-RA	9.40	462	470
CG10233-PB|CG10233-RB	9.40	462	470
CG10305-PA|RpS26-RA	10.04	1131	1139
CG10305-PB|RpS26-RB	10.04	1131	1139
CG10305-PC|RpS26-RC	10.04	1131	1139
CG10325-PB|abd-A-RB	9.67	449	457
CG10376-PA|CG10376-RA	9.67	370	378
CG10392-PA|Ogt-RA	9.40	1351	1359
CG10392-PB|Ogt-RB	9.40	1351	1359
CG10392-PC|Ogt-RC	9.40	1351	1359
CG10393-PA|amos-RA	9.23	1379	1387
CG10419-PA|CG10419-RA	11.21	959	967
CG10449-PA|Catsup-RA	9.96	194	202
CG10474-PA|CG10474-RA	9.06	1302	1310
CG10536-PA|cbx-RA	10.32	786	794
CG10536-PB|cbx-RB	10.32	945	953
CG10537-PA|Rdl-RA	10.94	829	837
CG10537-PB|Rdl-RB	10.94	829	837
CG10537-PC|Rdl-RC	10.94	829	837
CG10539-PA|S6k-RA	9.40	636	644
CG10541-PA|Tektin-C-RA	10.04	889	897
CG10588-PA|CG10588-RA	10.32	1101	1109
CG10593-PA|Acer-RA	9.06	231	239
CG10604-PA|bsh-RA	11.21	1468	1476
CG10605-PA|caup-RA	9.06	477	485
CG1063-PA|Itp-	9.96	839	847

r83A-RA			
CG1063-PB|Itp-r83A-RB	9.96	839	847
CG10654-PA|CG10654-RA	9.06	834	842
CG10674-PA|CG10674-RA	9.06	606	614
CG10692-PA|Dnmt2-RA	9.06	697	705
CG10692-PB|Dnmt2-RB	9.06	704	712
CG10704-PA|toe-RA	9.67	1069	1077
CG1071-PA|E2f2-RA	10.32	91	99
CG10741-PA|CG10741-RA	11.21	1054	1062
CG10757-PA|mRpS18B-RA	10.32	241	249
CG10757-PB|mRpS18B-RB	10.32	241	249
CG10778-PA|CG10778-RA	9.06	487	495
CG1078-PA|CG1078-RA	9.67	2	10
CG10794-PA|DptB-RA	10.94	1019	1027
CG10795-PA|CG10795-RA	9.23	404	412
CG10811-PA|eIF-4G-RA	9.23	1218	1226
CG10834-PA|CG10834-RA	11.21	1042	1050
CG10842-PA|Cyp4p1-RA	9.40	532	540
CG10844-PA|Rya-r44F-RA	10.32	1401	1409
CG10844-PB|Rya-r44F-RB	10.32	1401	1409
CG10844-PC|Rya-r44F-RC	10.32	1401	1409
CG10844-PD|Rya-r44F-RD	10.32	1401	1409
CG10859-PA|CG10859-RA	10.32	991	999
CG10864-PA|CG10864-RA	9.51	488	496
CG10887-PA|CG10887-RA	9.51	225	233
CG10915-PA|CG10915-RA	9.67	39	47
CG10932-PA|CG10932-RA	9.06	433	441
CG10949-	10.94	1170	1178

PA|CG10949-RA			
CG10954-PA|Arc-p34-RA	10.94	934	942
CG10989-PA|CG10989-RA	9.40	339	347
CG10997-PA|Clic-RA	9.06	257	265
CG11018-PA|CG11018-RA	9.23	883	891
CG11020-PA|nompC-RA	9.40	914	922
CG11020-PB|nompC-RB	9.40	914	922
CG11024-PA|cl-RA	10.94	753	761
CG11025-PA|isopeptidase-T-3-RA	9.51	1112	1120
CG11025-PB|isopeptidase-T-3-RB	9.51	600	608
CG11084-PA|pk-RA	9.67	467	475
CG1109-PA|CG1109-RA	9.96	117	125
CG1109-PB|CG1109-RB	9.96	117	125
CG11125-PA|CG11125-RA	9.96	716	724
CG11145-PA|CG11145-RA	11.21	602	610
CG11247-PA|CG11247-RA	9.40	632	640
CG11247-PB|CG11247-RB	9.40	632	640
CG11268-PA|CG11268-RA	9.40	324	332
CG11281-PA|CG11281-RA	9.40	184	192
CG11294-PA|CG11294-RA	10.94	351	359
CG11380-PA|CG11380-RA	9.67	1148	1156
CG11381-PA|CG11381-RA	9.96	84	92
CG11412-PA|CG11412-RA	9.51	501	509
CG11412-PB|CG11412-RB	9.51	501	509
CG11412-PC|CG11412-RC	9.51	501	509
CG1142-PA|CG1142-RA	9.23	282	290
CG11447-PA|CG11447-RA	9.06	534	542

CG11455-PA|CG11455-RA	9.23		1104		1112		
CG11455-PB|CG11455-RB	9.23		1104		1112		
CG11488-PA|mRpL10-RA	9.96		449		457		
CG11491-PA|br-RA	9.67		1463		1471		
CG11491-PB|br-RB	9.67		1463		1471		
CG11491-PC|br-RC	9.67		1463		1471		
CG11491-PD|br-RD	9.67		1463		1471		
CG11491-PE|br-RE	9.67		1463		1471		
CG11491-PF|br-RF	9.67		1463		1471		
CG11491-PG|br-RG	9.67		1463		1471		
CG11500-PA|CG11500-RA	9.06		302		310		
CG11502-PC|svp-RC	9.40		1366		1374		
CG11533-PB|CG11533-RB	9.40		1437		1445		
CG11539-PA|CG11539-RA	9.23		25		33		
CG11562-PA|CG11562-RA	11.21		1403		1411		
CG11567-PB|Cpr-RB	10.94		1025		1033		
CG11594-PA|CG11594-RA	9.40		1351		1359		
CG11594-PB|CG11594-RB	9.40		786		794		
CG11594-PC|CG11594-RC	9.40		1351		1359		
CG11604-PA|CG11604-RA	10.32		388		396		
CG11611-PA|Tim13-RA	10.94		54		62		
CG1171-PA|Akh-RA	11.21	9.40	511	209	519	217	
CG11734-PB|HERC2-RB	9.40		245		253		
CG11819-PA|CG11819-RA	9.06		1435		1443		
CG11837-PA|CG11837-RA	10.94		168		176		
CG11880-PA|CG11880-RA	9.06		8		16		
CG11880-PB|CG11880-RB	9.06		8		16		
CG11880-PC|CG11880-RC	9.06		8		16		
CG11895-PA|stan-RA	9.51		641		649		
CG11899-PA|CG11899-RA	9.06		1352		1360		
CG11908-PA|rha-RA	10.04		343		351		

CG1193-PB|CG1193-RB	11.21		814		822	
CG11940-PA|CG11940-RA	9.51		544		552	
CG11943-PA|CG11943-RA	9.51		1114		1122	
CG11943-PB|CG11943-RB	9.51		1114		1122	
CG12026-PA|CG12026-RA	9.67		1368		1376	
CG12026-PB|CG12026-RB	9.67		1368		1376	
CG12038-PA|CG12038-RA	9.67		430		438	
CG12038-PB|CG12038-RB	9.67		430		438	
CG12213-PA|CG12213-RA	9.96		996		1004	
CG12213-PB|CG12213-RB	9.96		996		1004	
CG12229-PA|CG12229-RA	10.94		1259		1267	
CG12231-PA|CG12231-RA	9.40		604		612	
CG12298-PA|sub-RA	9.67		195		203	
CG12310-PA|CG12310-RA	10.04		866		874	
CG12347-PA|CG12347-RA	9.40		114		122	
CG12352-PA|san-RA	9.67	9.06	1396	346	1404	354
CG12370-PA|CG12370-RA	10.04		1139		1147	
CG12370-PB|CG12370-RB	10.04		1139		1147	
CG12384-PA|CG12384-RA	9.06		826		834	
CG12399-PA|Mad-RA	10.32		1326		1334	
CG12402-PA|CG12402-RA	11.21		985		993	
CG12449-PB|CG12449-RB	10.04		385		393	
CG12449-PC|CG12449-RC	10.04		385		393	
CG12449-PE|CG12449-RE	10.04		385		393	
CG12449-PF|CG12449-RF	10.04		385		393	
CG12488-PA|CG12488-RA	9.96		1059		1067	
CG12502-PA|CG12502-RA	9.40		461		469	

CG12538-PA|CG12538-RA	9.23	25	33
CG12564-PA|CG12564-RA	9.51	1243	1251
CG12605-PC|CG12605-RC	11.21	1053	1061
CG12609-PA|CG12609-RA	9.40	1422	1430
CG12609-PB|CG12609-RB	9.40	1422	1430
CG12621-PA|beat-IIIa-RA	9.23	258	266
CG12632-PB|fd3F-RB	9.51	1330	1338
CG12673-PA|olf413-RA	9.40	1190	1198
CG12714-PB|CG12714-RB	9.40	229	237
CG12715-PA|CG12715-RA	10.94	401	409
CG12767-PA|Dip3-RA	9.40	1450	1458
CG12787-PA|hoe1-RA	11.21	162	170
CG12795-PA|CG12795-RA	9.96	1481	1489
CG12833-PA|esn-RA	10.32	171	179
CG12833-PB|esn-RB	10.32	171	179
CG12873-PA|CG12873-RA	9.96	624	632
CG12906-PA|Gr47a-RA	10.32	1331	1339
CG12913-PA|CG12913-RA	10.04	364	372
CG12952-PA|sage-RA	11.21	472	480
CG12952-PB|sage-RB	11.21	202	210
CG13002-PA|CG13002-RA	9.67	857	865
CG13047-PA|CG13047-RA	9.06	778	786
CG13057-PA|retinin-RA	11.21	550	558
CG13068-PA|CG13068-RA	9.06	52	60
CG13082-PA|CG13082-RA	10.04	799	807
CG13088-PA|CG13088-RA	10.04	615	623
CG13111-PA|CG13111-RA	9.06	661	669
CG13121-PA|CG13121-RA	9.06	113	121
CG13130-PA|CG13130-RA	9.40	887	895

CG13142-PA|CG13142-RA	9.40	59	67
CG13171-PA|CG13171-RA	10.32	278	286
CG1318-PC|Hexo1-RC	10.04	1303	1311
CG13201-PA|ix-RA	9.67	858	866
CG13239-PA|CG13239-RA	9.67	449	457
CG13298-PA|CG13298-RA	9.06	361	369
CG13308-PA|CG13308-RA	9.67	157	165
CG13312-PA|CG13312-RA	9.67	1491	1499
CG13313-PA|CG13313-RA	9.51	929	937
CG13315-PA|CG13315-RA	9.67	1191	1199
CG13361-PA|CG13361-RA	9.67	1230	1238
CG13377-PA|CG13377-RA	10.94	594	602
CG1338-PA|hydra-RA	9.23	1204	1212
CG1341-PA|Rpt1-RA	9.96	589	597
CG13436-PA|CG13436-RA	10.04	219	227
CG13455-PA|CG13455-RA	10.04	1441	1449
CG13457-PA|CG13457-RA	9.23	1293	1301
CG13479-PA|CG13479-RA	9.40	835	843
CG13504-PA|CG13504-RA	9.40	802	810
CG13524-PA|Obp58c-RA	10.04	1112	1120
CG13545-PA|CG13545-RA	10.04	278	286
CG13567-PA|CG13567-RA	9.40	878	886
CG13567-PB|CG13567-RB	9.40	878	886
CG13581-PA|CG13581-RA	11.21	1147	1155
CG13585-PA|CG13585-RA	9.40	368	376
CG13585-PB|CG13585-RB	9.40	368	376
CG13592-PA|CG13592-RA	9.40	1353	1361
CG13618-	10.04	1362	1370

PA|CG13618-RA			
CG13622-PA|CG13622-RA	10.04	67	75
CG13659-PA|CG13659-RA	9.23	1338	1346
CG13738-PA|CG13738-RA	9.40	682	690
CG13743-PA|CG13743-RA	9.67	1483	1491
CG13790-PA|CG13790-RA	9.40	1205	1213
CG13802-PA|CG13802-RA	10.04	1189	1197
CG13877-PA|CG13877-RA	9.67	1417	1425
CG13885-PA|CG13885-RA	11.21	758	766
CG13897-PA|CG13897-RA	9.51	1169	1177
CG13969-PA|bwa-RA	9.06	295	303
CG14014-PA|CG14014-RA	9.40	103	111
CG14021-PA|CG14021-RA	9.23	1308	1316
CG14021-PB|CG14021-RB	9.23	1308	1316
CG14034-PA|CG14034-RA	9.96	1293	1301
CG14059-PA|CG14059-RA	9.23	637	645
CG14073-PA|CG14073-RA	10.94	596	604
CG14073-PB|CG14073-RB	9.67	431	439
CG14104-PA|CG14104-RA	9.96	1346	1354
CG14110-PA|CG14110-RA	11.21	773	781
CG14131-PA|CG14131-RA	10.94	130	138
CG14141-PA|CG14141-RA	9.23	233	241
CG14164-PA|CG14164-RA	9.67	340	348
CG14222-PA|CG14222-RA	9.40	1041	1049
CG14225-PA|CG14225-RA	9.06	1040	1048
CG14229-PA|CG14229-RA	10.32	20	28
CG14249-PA|beat-VII-RA	10.04	636	644

CG14297-PA|CG14297-RA	9.40		28		36	
CG14317-PA|CG14317-RA	9.23		1257		1265	
CG14318-PA|CG14318-RA	9.51		81		89	
CG14329-PA|CG14329-RA	11.21		1069		1077	
CG14360-PA|Or88a-RA	11.21		379		387	
CG14374-PA|CG14374-RA	9.96		615		623	
CG14387-PA|CG14387-RA	9.06		1170		1178	
CG14489-PA|olf186-M-RA	11.21		1392		1400	
CG14506-PA|CG14506-RA	10.04	9.40	951	917	959	925
CG14540-PA|CG14540-RA	10.94		622		630	
CG14542-PA|CG14542-RA	10.04		406		414	
CG14561-PA|CG14561-RA	9.23		1374		1382	
CG14565-PA|CG14565-RA	10.32		785		793	
CG14573-PA|CG14573-RA	9.40		566		574	
CG14659-PA|CG14659-RA	10.32		366		374	
CG14711-PA|CG14711-RA	9.40		202		210	
CG14720-PA|CG14720-RA	9.96		929		937	
CG14731-PA|CG14731-RA	9.96	9.23	665	1423	673	1431
CG14744-PA|CG14744-RA	10.04		1149		1157	
CG14746-PA|PGRP-SC1a-RA	10.04		1209		1217	
CG14778-PA|CG14778-RA	9.23		300		308	
CG1480-PA|bnk-RA	9.67		1253		1261	
CG14949-PA|CG14949-RA	9.06		1017		1025	
CG15020-PA|CG15020-RA	10.32		1488		1496	
CG15096-PA|CG15096-RA	9.06		1083		1091	
CG15111-PB|CG15111-RB	10.32		1143		1151	
CG1513-PA|CG1513-	9.67		502		510	

RA									
CG1515-PA|l(1)G0155-RA	9.67			283			291		
CG15166-PA|CG15166-RA	9.40			854			862		
CG15199-PA|CG15199-RA	9.06			1461			1469		
CG15233-PA|CG15233-RA	9.06			723			731		
CG15234-PA|CG15234-RA	11.21			186			194		
CG15259-PA|nht-RA	9.23			1426			1434		
CG1527-PA|RpS14b-RA	9.06			497			505		
CG15288-PA|wb-RA	9.23			252			260		
CG15288-PB|wb-RB	9.23			252			260		
CG15343-PA|CG15343-RA	9.06			523			531		
CG15373-PA|CG15373-RA	10.32	9.40	9.23	1049	1244	253	1057	1252	261
CG15393-PA|CG15393-RA	9.51	9.06		650	947		658	955	
CG15418-PA|CG15418-RA	9.67			725			733		
CG15488-PA|CG15488-RA	9.40			121			129		
CG15553-PA|CG15553-RA	9.23			1324			1332		
CG15589-PA|CG15589-RA	9.40			1277			1285		
CG15594-PA|CG15594-RA	9.96			970			978		
CG15594-PB|CG15594-RB	9.96			922			930		
CG15604-PA|CG15604-RA	11.21			938			946		
CG15671-PA|cv-2-RA	10.32	9.67	9.23	411	832	810	419	840	818
CG15719-PA|CG15719-RA	10.32			1041			1049		
CG15811-PA|Rop-RA	9.40			13			21		
CG15817-PA|CG15817-RA	10.32			1324			1332		
CG15873-PA|CG15873-RA	9.23			64			72		
CG15897-PA|CG15897-RA	10.04			571			579		
CG1616-PA|dpa-RA	10.32			433			441		
CG1624-PA|dpld-RA	9.67			1084			1092		
CG1624-PB|dpld-RB	9.67			1084			1092		
CG1624-PC|dpld-RC	9.67			1084			1092		

CG1659-PA|unc-119-RA	9.40		944		952	
CG1669-PA|kappaB-Ras-RA	9.96		319		327	
CG16723-PA|CG16723-RA	9.40		384		392	
CG16743-PA|CG16743-RA	9.67		557		565	
CG16793-PA|CG16793-RA	9.06		951		959	
CG16813-PA|CG16813-RA	9.23		1253		1261	
CG16833-PA|CG16833-RA	9.40		615		623	
CG16885-PA|CG16885-RA	10.94		142		150	
CG16885-PB|CG16885-RB	10.94		142		150	
CG1698-PA|CG1698-RA	9.96		1419		1427	
CG16985-PA|CG16985-RA	9.23		614		622	
CG17031-PA|ref2-RA	9.40		465		473	
CG17065-PA|CG17065-RA	10.04	9.06	762	994	770	1002
CG17129-PA|CG17129-RA	9.23		961		969	
CG17129-PB|CG17129-RB	9.23		961		969	
CG17129-PC|CG17129-RC	9.23		961		969	
CG17152-PA|CG17152-RA	9.96		863		871	
CG17309-PB|Csk-RB	9.40		44		52	
CG17309-PC|Csk-RC	9.40		44		52	
CG17309-PD|Csk-RD	9.40		44		52	
CG17309-PE|Csk-RE	9.40		44		52	
CG17386-PA|CG17386-RA	10.94		1220		1228	
CG1745-PA|CG1745-RA	9.23		1277		1285	
CG1745-PB|CG1745-RB	9.23		1277		1285	
CG17461-PA|Kif3C-RA	10.94		558		566	
CG17592-PA|Usf-RA	9.06		718		726	
CG17592-PB|Usf-RB	9.06		718		726	
CG17652-PA|CG17652-RA	9.96		872		880	
CG17707-PA|CG17707-RA	9.96		31		39	

CG17726-PA|CG17726-RA	9.96	149	157
CG1773-PA|CG1773-RA	10.32	817	825
CG17735-PA|CG17735-RA	11.21	1111	1119
CG17824-PA|CG17824-RA	9.40	852	860
CG17838-PD|CG17838-RD	10.04	612	620
CG17838-PE|CG17838-RE	10.04	612	620
CG17841-PA|CG17841-RA	9.40	252	260
CG17888-PF|Pdp1-RF	9.40	990	998
CG17927-PA|Mhc-RA	9.96	1147	1155
CG17927-PB|Mhc-RB	9.96	1147	1155
CG17927-PC|Mhc-RC	9.96	1147	1155
CG17927-PD|Mhc-RD	9.96	1147	1155
CG17927-PE|Mhc-RE	9.96	1147	1155
CG17927-PF|Mhc-RF	9.96	1147	1155
CG17927-PG|Mhc-RG	9.96	1147	1155
CG17927-PH|Mhc-RH	9.96	1147	1155
CG17927-PI|Mhc-RI	9.96	1147	1155
CG17927-PJ|Mhc-RJ	9.96	1147	1155
CG17927-PK|Mhc-RK	9.96	1147	1155
CG17927-PL|Mhc-RL	9.96	1147	1155
CG17927-PM|Mhc-RM	9.96	1147	1155
CG1794-PA|Mmp2-RA	9.51	784	792
CG1794-PB|Mmp2-RB	9.51	784	792
CG17958-PA|Sry-delta-RA	9.67	703	711
CG17991-PA|CG17991-RA	9.06	39	47
CG1800-PA|pasha-RA	9.23	434	442
CG18003-PA|CG18003-RA	9.40	407	415
CG18003-PB|CG18003-RB	9.40	305	313
CG18023-PB|Eip78C-RB	10.04	929	937
CG18023-PC|Eip78C-RC	10.04	767	775
CG18063-PA|CG18063-RA	9.96	434	442
CG18063-PB|CG18063-RB	9.96	434	442
CG18064-PA|Met75Cb-RA	9.40	18	26

CG18087-PA|Sgs7-RA	9.40	283	291
CG18110-PA|CG18110-RA	9.06	1221	1229
CG18347-PA|CG18347-RA	9.96	127	135
CG18396-PA|Mst98Cb-RA	10.32	935	943
CG18408-PB|Cap-RB	9.06	364	372
CG18408-PG|CAP-RG	9.06	364	372
CG18408-PH|CAP-RH	9.06	364	372
CG18408-PJ|CAP-RJ	9.06	733	741
CG18417-PA|CG18417-RA	10.32	1050	1058
CG1842-PA|Dhc98D-RA	9.96	766	774
CG1849-PA|run-RA	9.40	376	384
CG18519-PA|CG18519-RA	9.06	325	333
CG18519-PB|CG18519-RB	9.06	325	333
CG18540-PA|CG18540-RA	10.32	613	621
CG18554-PA|CG18554-RA	9.40	197	205
CG18581-PA|CG18581-RA	10.04	300	308
CG18606-PA|CG18606-RA	9.06	268	276
CG1864-PC|Hr38-RC	9.40	971	979
CG18660-PA|Nckx30C-RA	9.40	1197	1205
CG18660-PB|Nckx30C-RB	9.40	1197	1205
CG18660-PC|Nckx30C-RC	9.40	1197	1205
CG18679-PA|CG18679-RA	9.67	156	164
CG18731-PA|CG18731-RA	9.23	1416	1424
CG1886-PA|ATP7-RA	10.32	457	465
CG1893-PA|CG1893-RA	9.06	376	384
CG1942-PA|CG1942-RA	9.40	264	272
CG1969-PA|CG1969-RA	9.96	24	32
CG1969-PB|CG1969-RB	9.96	452	460
CG2040-PA|hig-RA	9.67	1433	1441
CG2040-PB|hig-RB	9.67	1433	1441

CG2040-PC|hig-RC	9.67			1433		1441	
CG2040-PD|hig-RD	9.67			1433		1441	
CG2044-PA|Lcp4-RA	11.21			1210		1218	
CG2054-PA|Cht2-RA	10.32			387		395	
CG2056-PA|CG2056-RA	9.96			205		213	
CG2056-PB|CG2056-RB	9.96			205		213	
CG2076-PA|CG2076-RA	9.51			612		620	
CG2098-PB|ferrochelatase-RB	9.40			768		776	
CG2182-PB|CG2182-RB	9.96			620		628	
CG2221-PA|l(1)G0289-RA	9.40			597		605	
CG2381-PF|Syt7-RF	9.06			1135		1143	
CG2525-PA|Hus1-like-RA	9.40			877		885	
CG2789-PA|CG2789-RA	9.06			98		106	
CG2849-PA|Rala-RA	9.23			306		314	
CG2849-PB|Rala-RB	9.23			306		314	
CG2849-PC|Rala-RC	9.23			306		314	
CG2879-PA|CG2879-RA	11.21	9.96		761	869	769	877
CG2934-PA|VhaAC39-RA	10.04			1217		1225	
CG2979-PA|Yp2-RA	9.51			1278		1286	
CG2980-PA|thoc5-RA	11.21			905		913	
CG2985-PA|Yp1-RA	9.51			433		441	
CG30030-PA|Gr47b-RA	9.67			321		329	
CG30036-PA|CG30036-RA	9.06			1049		1057	
CG30048-PA|CG30048-RA	9.23			814		822	
CG30049-PA|CG30049-RA	10.32			669		677	
CG30061-PA|CG30061-RA	9.51			641		649	
CG30077-PA|CG30077-RA	9.06			244		252	
CG30087-PA|CG30087-RA	9.06			1052		1060	
CG30094-PA|CG30094-RA	9.06			756		764	
CG30169-PA|CG30169-RA	10.04			534		542	

CG30268-PA|CG30268-RA	9.40	726	734
CG30275-PA|CG30275-RA	10.04	718	726
CG30275-PC|CG30275-RC	10.04	718	726
CG30284-PA|CG30284-RA	9.51	1361	1369
CG30284-PB|CG30284-RB	9.51	1361	1369
CG30342-PA|CG30342-RA	10.32	1267	1275
CG30360-PA|CG30360-RA	10.04	819	827
CG30360-PB|CG30360-RB	10.04	819	827
CG30464-PA|CG30464-RA	9.06	391	399
CG30468-PA|CG30468-RA	9.96	485	493
CG30471-PA|CG30471-RA	10.32	812	820
CG30490-PA|CG30490-RA	11.21	739	747
CG30491-PA|CG30491-RA	9.96	822	830
CG3060-PA|mr-RA	10.32	972	980
CG3065-PA|CG3065-RA	10.94	38	46
CG3065-PB|CG3065-RB	10.94	38	46
CG31036-PA|CG31036-RA	10.32	187	195
CG31038-PC|CG31038-RC	9.40	1317	1325
CG31051-PA|CG31051-RA	9.06	46	54
CG31064-PA|CG31064-RA	9.06	523	531
CG31064-PB|CG31064-RB	9.06	523	531
CG31064-PE|CG31064-RE	9.96	492	500
CG31078-PA|CG31078-RA	9.51	379	387
CG31094-PA|LpR1-RA	10.32	1353	1361
CG31098-PA|CG31098-RA	10.04	1468	1476
CG31104-PA|CG31104-RA	9.40	274	282
CG31111-PA|CG31111-RA	9.06	556	564

CG31118-PA|RabX4-RA	9.06	733	741
CG31122-PA|CG31122-RA	9.40	978	986
CG31131-PA|CG31131-RA	9.67	872	880
CG31152-PA|CG31152-RA	9.40	908	916
CG31157-PA|CG31157-RA	10.04	1025	1033
CG31174-PA|CG31174-RA	9.96	117	125
CG3121-PA|CG3121-RA	9.51	651	659
CG31253-PA|CG31253-RA	9.67	899	907
CG31288-PA|CG31288-RA	9.23	245	253
CG31327-PA|CG31327-RA	9.40	627	635
CG31367-PA|CG31367-RA	10.04	623	631
CG31370-PA|CG31370-RA	9.40	459	467
CG31371-PA|CG31371-RA	9.06	555	563
CG31431-PA|CG31431-RA	9.96	356	364
CG31523-PA|CG31523-RA	9.40	346	354
CG31523-PB|CG31523-RB	9.40	346	354
CG31523-PC|CG31523-RC	9.40	346	354
CG31523-PD|CG31523-RD	9.40	346	354
CG31536-PA|Cdep-RA	9.40	63	71
CG31536-PB|Cdep-RB	9.40	63	71
CG31536-PC|Cdep-RC	9.40	63	71
CG3161-PA|Vha16-RA	11.21	563	571
CG3161-PB|Vha16-RB	11.21	563	571
CG3161-PC|Vha16-RC	11.21	563	571
CG3161-PD|Vha16-RD	11.21	563	571
CG31612-PA|CG31612-RA	10.94	804	812
CG31622-PA|Gr39a-RA	9.96	463	471
CG31623-PA|dtr-RA	9.96	201	209
CG31645-PA|CG31645-RA	9.23	1384	1392

CG3169-PA|Spt3-RA	9.23		1310		1318	
CG31690-PB|CG31690-RB	9.67		422		430	
CG31699-PA|CG31699-RA	9.06		722		730	
CG31731-PA|CG31731-RA	9.06		442		450	
CG31732-PE|yuri-RE	10.04		1232		1240	
CG31732-PG|yuri-RG	10.04		1232		1240	
CG31752-PA|CG31752-RA	10.04		652		660	
CG31755-PA|CG31755-RA	10.32		567		575	
CG31777-PA|CG31777-RA	10.94		301		309	
CG31790-PA|CG31790-RA	9.40	9.40	135	881	143	889
CG31867-PA|CG31867-RA	9.06		357		365	
CG31871-PA|CG31871-RA	11.21		292		300	
CG31929-PA|Gr22c-RA	9.06		1140		1148	
CG31957-PA|CG31957-RA	9.40		510		518	
CG31959-PB|CG31959-RB	11.21		923		931	
CG31973-PA|CG31973-RA	10.32		482		490	
CG32024-PA|CG32024-RA	9.67		1076		1084	
CG32026-PA|CG32026-RA	10.32		1417		1425	
CG32079-PA|CG32079-RA	11.21		1378		1386	
CG32231-PA|CG32231-RA	10.04		102		110	
CG32258-PA|Gr64e-RA	11.21		1124		1132	
CG32258-PB|Gr64e-RB	11.21		1124		1132	
CG32261-PA|Gr64a-RA	9.40		4		12	
CG32319-PA|CG32319-RA	9.40		356		364	
CG32333-PA|CG32333-RA	10.32		579		587	
CG32333-PB|CG32333-RB	10.32		579		587	
CG32452-PA|CG32452-RA	10.04		1141		1149	

CG32452-PB|CG32452-RB	10.04		1141		1149	
CG32466-PB|rn-RB	10.94		1090		1098	
CG32490-PJ|cpx-RJ	9.67		401		409	
CG32490-PK|cpx-RK	9.67		401		409	
CG32508-PA|shakB-RA	9.67		1182		1190	
CG32508-PB|shakB-RB	9.67		1182		1190	
CG32548-PC|CG32548-RC	9.96	9.06	128	1285	136	1293
CG32551-PA|CG32551-RA	9.96		1097		1105	
CG32564-PA|CG32564-RA	9.40		1452		1460	
CG32667-PA|CG32667-RA	9.06		674		682	
CG32788-PB|Crg-1-RB	9.51		1330		1338	
CG3279-PA|CG3279-RA	10.32		465		473	
CG32798-PA|CG32798-RA	11.21		1448		1456	
CG32812-PA|CG32812-RA	10.04		962		970	
CG32817-PA|CG32817-RA	10.94		137		145	
CG32825-PA|Or19b-RA	9.51		594		602	
CG32835-PA|CG32835-RA	9.40		1		9	
CG33056-PA|CG33056-RA	9.23		584		592	
CG33056-PB|CG33056-RB	9.23		584		592	
CG33056-PC|CG33056-RC	9.23		584		592	
CG33056-PD|CG33056-RD	9.23		584		592	
CG33056-PE|CG33056-RE	9.23		584		592	
CG33060-PA|CG33060-RA	11.21		1490		1498	
CG33080-PA|CG33080-RA	9.23		729		737	
CG33113-PC|Rtnl1-RC	9.51		467		475	
CG3313-PA|CG3313-RA	9.23		1276		1284	
CG33131-PA|SCAP-RA	9.67		1289		1297	
CG33133-PA|grau-RA	9.40		462		470	

CG33136-PA|CG33136-RA	9.40		884		892		
CG33162-PA|SrpRbeta-RA	9.40		1151		1159		
CG33169-PA|CG33169-RA	11.21	9.67	1274	161	1282	169	
CG33169-PB|CG33169-RB	11.21	9.67	1274	161	1282	169	
CG33174-PA|CG33174-RA	9.67		575		583		
CG33174-PD|CG33174-RD	9.67		575		583		
CG33181-PA|CG33181-RA	9.06		727		735		
CG33226-PA|CG33226-RA	9.23		737		745		
CG33232-PB|CG33232-RB	9.06		25		33		
CG33232-PC|CG33232-RC	9.06		25		33		
CG33232-PD|CG33232-RD	9.06		25		33		
CG3327-PB|E23-RB	9.23		1143		1151		
CG33306-PA|CG33306-RA	9.40		490		498		
CG33311-PA|CG33311-RA	9.40		553		561		
CG33335-PA|CG33335-RA	9.51	9.40	586	811	594	819	
CG33337-PA|CG33337-RA	10.94		782		790		
CG33530-PA|Acp53C14c-RA	9.67		1248		1256		
CG33542-PA|upd3-RA	9.51		1442		1450		
CG33639-PA|CG33639-RA	9.67		146		154		
CG33692-PA|CG33692-RA	9.51		521		529		
CG33692-PB|CG33692-RB	9.51		521		529		
CG33692-PC|CG33692-RC	9.51		521		529		
CG33757-PA|CG33757-RA	9.51		789		797		
CG33957-PC|cp309-RC	9.40		702		710		
CG33995-PA|CG33995-RA	9.51		1331		1339		
CG33995-PB|CG33995-RB	9.51		1331		1339		
CG33995-PC|CG33995-RC	9.51		1331		1339		

CG34001-PA|CG34001-RA	9.06	482	490
CG34016-PA|CG34016-RA	9.67	92	100
CG34043-PA|CG34043-RA	10.32	92	100
CG3423-PA|sa-RA	9.51	226	234
CG3431-PA|Uch-L3-RA	9.67	1325	1333
CG3436-PA|CG3436-RA	9.23	435	443
CG3436-PB|CG3436-RB	9.23	435	443
CG3484-PA|Adhr-RA	9.40	1032	1040
CG3484-PB|Adhr-RB	9.40	1032	1040
CG3523-PA|CG3523-RA	9.40	1374	1382
CG3565-PA|CG3565-RA	9.96	736	744
CG3604-PA|CG3604-RA	9.40	214	222
CG3626-PA|CG3626-RA	9.06	1420	1428
CG3654-PD|CG3654-RD	9.67	1009	1017
CG3679-PA|CG3679-RA	9.23	1372	1380
CG3753-PA|Marcal1-RA	9.06	400	408
CG3779-PA|numb-RA	10.94	163	171
CG3805-PA|CG3805-RA	9.06	1341	1349
CG3845-PA|l(2)01424-RA	9.96	829	837
CG3845-PB|l(2)01424-RB	9.96	829	837
CG3894-PA|CG3894-RA	9.40	906	914
CG3894-PB|CG3894-RB	9.40	906	914
CG3915-PB|Drl-2-RB	11.21	228	236
CG3924-PA|Chi-RA	9.06	255	263
CG3924-PB|Chi-RB	9.06	418	426
CG3927-PA|CG3927-RA	10.04	759	767
CG3937-PB|cher-RB	10.04	178	186
CG3937-PC|cher-RC	10.04	178	186
CG3986-PA|Cht4-RA	9.06	348	356
CG3988-PA|gammaSnap-RA	9.40	702	710

CG40000-PA|CG40000-RA	9.67	986	994
CG40068-PA|CG40068-RA	9.06	886	894
CG40131-PA|CG40131-RA	9.51	286	294
CG4017-PA|CG4017-RA	9.67	1134	1142
CG40195-PA|CG40195-RA	9.23	954	962
CG40228-PA|CG40228-RA	9.06	348	356
CG4029-PA|jumu-RA	9.51	1060	1068
CG40306-PB|CG40306-RB	9.67	691	699
CG40351-PA|CG40351-RA	9.23	506	514
CG40351-PB|CG40351-RB	9.23	506	514
CG40409-PA|CG40409-RA	9.67	250	258
CG4076-PA|Nufip-RA	9.06	161	169
CG4084-PA|l(2)not-RA	11.21	271	279
CG4095-PA|CG4095-RA	9.23	1300	1308
CG41061-PA|CG41061-RA	10.94	71	79
CG41065-PA|CG41065-RA	11.21	198	206
CG41107-PA|CG41107-RA	10.04	94	102
CG41138-PA|CG41138-RA	9.96	552	560
CG4199-PA|CG4199-RA	9.23	75	83
CG4290-PA|CG4290-RA	9.96	470	478
CG4306-PA|CG4306-RA	9.67	90	98
CG4307-PA|Oscp-RA	11.21	295	303
CG4479-PA|Mst35Ba-RA	9.67	1319	1327
CG4502-PA|CG4502-RA	10.94	155	163
CG4502-PB|CG4502-RB	10.94	155	163
CG4553-PA|CG4553-RA	10.94	1219	1227
CG4573-PA|CG4573-RA	9.40	760	768

CG4587-PA|CG4587-RA	9.96	540	548
CG4591-PA|Tsp86D-RA	10.04	523	531
CG4655-PB|CG4655-RB	9.23	224	232
CG4678-PA|CG4678-RA	9.06	806	814
CG4678-PB|CG4678-RB	9.06	806	814
CG4678-PC|CG4678-RC	9.06	806	814
CG4722-PA|bib-RA	9.40	1205	1213
CG4747-PA|CG4747-RA	10.32	601	609
CG4830-PA|CG4830-RA	9.96	383	391
CG4852-PA|Sras-RA	9.40	1017	1025
CG4877-PA|CG4877-RA	9.40	842	850
CG4877-PB|CG4877-RB	9.40	842	850
CG4940-PA|CG4940-RA	9.23	1301	1309
CG4989-PA|CG4989-RA	9.96	635	643
CG5002-PA|CG5002-RA	9.67	342	350
CG5012-PA|mRpL12-RA	10.32	1420	1428
CG5050-PA|CG5050-RA	9.40	1193	1201
CG5106-PA|scpr-C-RA	9.96	1413	1421
CG5137-PA|Cyp312a1-RA	9.06	1396	1404
CG5151-PA|CG5151-RA	10.94	1216	1224
CG5162-PA|CG5162-RA	9.96	546	554
CG5179-PA|Cdk9-RA	9.23	918	926
CG5220-PA|CG5220-RA	10.94	661	669
CG5241-PA|CG5241-RA	9.67	815	823
CG5246-PA|CG5246-RA	10.94	1255	1263
CG5248-PA|loco-RA	9.06	226	234
CG5279-PA|Rh5-RA	10.94	1270	1278
CG5319-PA|CG5319-RA	9.51	128	136

CG5374-PA|T-cp1-RA	9.23		1157		1165	
CG5374-PB|T-cp1-RB	9.23		1157		1165	
CG5424-PC|f-RC	10.04		1395		1403	
CG5428-PA|CG5428-RA	9.96		1143		1151	
CG5446-PA|CG5446-RA	9.67		378		386	
CG5481-PA|lea-RA	11.21		1296		1304	
CG5501-PB|Myo95E-RB	9.40		627		635	
CG5501-PD|Myo95E-RD	9.40		627		635	
CG5501-PE|Myo95E-RE	9.40		627		635	
CG5501-PF|Myo95E-RF	9.40		627		635	
CG5501-PG|Myo95E-RG	9.40		627		635	
CG5507-PA|T48-RA	9.96		702		710	
CG5529-PA|B-H1-RA	10.04		206		214	
CG5549-PA|CG5549-RA	10.04		1311		1319	
CG5580-PB|sbb-RB	10.32		58		66	
CG5603-PA|CYLD-RA	9.40		1231		1239	
CG5603-PB|CYLD-RB	9.40		155		163	
CG5603-PC|CYLD-RC	9.40		910		918	
CG5603-PD|CYLD-RD	9.40		1231		1239	
CG5603-PE|CYLD-RE	9.40		1231		1239	
CG5638-PA|Rh7-RA	9.40		689		697	
CG5683-PA|Aef1-RA	9.23		550		558	
CG5683-PB|Aef1-RB	9.23		550		558	
CG5683-PC|Aef1-RC	9.23		550		558	
CG5684-PA|CG5684-RA	9.96		381		389	
CG5740-PA|CG5740-RA	9.06		552		560	
CG5780-PA|CG5780-RA	9.40		142		150	
CG5785-PB|thr-RB	9.67		908		916	
CG5798-PA|CG5798-RA	9.40		212		220	
CG5819-PA|CG5819-RA	10.04	9.96	813	393	821	401
CG5819-PB|CG5819-RB	10.04	9.96	813	393	821	401
CG5847-PA|CG5847-RA	10.32		1089		1097	
CG5860-PA|CG5860-RA	9.06		729		737	

CG5877-PB|CG5877-RB	9.40		1476		1484	
CG5903-PA|CG5903-RA	9.23		253		261	
CG5907-PA|Frq2-RA	10.32		1282		1290	
CG5907-PB|Frq2-RB	10.32		1282		1290	
CG5907-PC|Frq2-RC	10.32		1282		1290	
CG5911-PA|ETHR-RA	10.32		209		217	
CG5911-PB|ETHR-RB	10.32		209		217	
CG5923-PA|DNApol-alpha73-RA	9.96		1183		1191	
CG5923-PB|DNApol-alpha73-RB	9.96		1183		1191	
CG5945-PA|CG5945-RA	9.67		914		922	
CG5987-PA|CG5987-RA	10.04		759		767	
CG5989-PA|CG5989-RA	9.51		390		398	
CG6005-PA|CG6005-RA	9.06		138		146	
CG6027-PA|cdi-RA	9.96		1463		1471	
CG6030-PA|ATPsyn-d-RA	9.67		783		791	
CG6030-PB|ATPsyn-d-RB	9.67		783		791	
CG6048-PA|CG6048-RA	9.51		783		791	
CG6074-PA|CG6074-RA	10.32		1435		1443	
CG6106-PA|CG6106-RA	11.21		1157		1165	
CG6118-PA|CG6118-RA	9.51		1055		1063	
CG6265-PA|Nep5-RA	9.40		630		638	
CG6265-PB|Nep5-RB	9.40		630		638	
CG6290-PA|CG6290-RA	9.06		1332		1340	
CG6324-PA|CG6324-RA	9.96	9.40	249	4	257	12
CG6327-PD|CG6327-RD	11.21		72		80	
CG6447-PA|CG6447-RA	9.67		259		267	
CG6447-PB|CG6447-RB	9.67		177		185	
CG6451-PA|blue-RA	9.06		841		849	
CG6452-PA|CG6452-RA	9.67		1354		1362	
CG6455-PA|CG6455-	9.06		247		255	

RA						
CG6476-PA|Su(var)3-9-RA	11.21		385		393	
CG6476-PB|Su(var)3-9-RB	11.21		385		393	
CG6511-PA|CG6511-RA	9.40		397		405	
CG6521-PA|Stam-RA	10.04		75		83	
CG6525-PA|CG6525-RA	11.21	10.32	698	150	706	158
CG6584-PA|SelR-RA	10.04		358		366	
CG6584-PB|SelR-RB	10.04		358		366	
CG6584-PC|SelR-RC	10.04		1479		1487	
CG6584-PD|SelR-RD	10.04		358		366	
CG6584-PE|SelR-RE	10.04		1479		1487	
CG6584-PF|SelR-RF	10.04		358		366	
CG6627-PA|Dnz1-RA	10.04		362		370	
CG6643-PA|CG6643-RA	9.96		324		332	
CG6659-PA|CG6659-RA	10.32		416		424	
CG6675-PA|CG6675-RA	11.21		80		88	
CG6685-PA|CG6685-RA	9.06		930		938	
CG6701-PA|CG6701-RA	9.23		934		942	
CG6753-PA|CG6753-RA	11.21	9.23	1007	1003	1015	1011
CG6806-PA|Lsp2-RA	9.96		115		123	
CG6833-PA|CG6833-RA	9.06		486		494	
CG6838-PA|CG6838-RA	10.04		926		934	
CG6838-PB|CG6838-RB	10.04		926		934	
CG6847-PA|CG6847-RA	9.40		1007		1015	
CG6885-PA|CG6885-RA	11.21		142		150	
CG6963-PA|gish-RA	10.32		395		403	
CG6963-PC|gish-RC	10.32		395		403	
CG6963-PG|gish-RG	10.32		799		807	
CG6990-PA|HP1c-RA	9.51	9.06	894	830	902	838
CG7045-PA|CG7045-RA	9.23		1241		1249	
CG7095-PA|CG7095-RA	9.51		1203		1211	
CG7101-PA|CG7101-	9.23		180		188	

RA			
CG7110-PB|CG7110-RB	10.32	1204	1212
CG7115-PA|CG7115-RA	9.40	416	424
CG7115-PB|CG7115-RB	9.40	416	424
CG7139-PA|CG7139-RA	9.51	779	787
CG7149-PA|CG7149-RA	9.67	249	257
CG7172-PA|CG7172-RA	9.51	180	188
CG7252-PA|CG7252-RA	11.21	597	605
CG7274-PA|CG7274-RA	10.04	948	956
CG7332-PA|CG7332-RA	10.32	1296	1304
CG7334-PA|Sug-RA	11.21	385	393
CG7334-PB|Sug-RB	11.21	385	393
CG7381-PA|CG7381-RA	9.06	1255	1263
CG7427-PA|CG7427-RA	9.40	658	666
CG7428-PA|halo-RA	9.96	705	713
CG7454-PA|Or85a-RA	10.04	1455	1463
CG7456-PA|CG7456-RA	9.40	157	165
CG7457-PA|CG7457-RA	11.21	624	632
CG7462-PB|Ank2-RB	9.06	860	868
CG7462-PC|Ank2-RC	9.06	860	868
CG7484-PB|CG7484-RB	9.67	1377	1385
CG7523-PA|CG7523-RA	9.67	931	939
CG7573-PB|CG7573-RB	9.96	1152	1160
CG7577-PA|ppk20-RA	9.06	1492	1500
CG7592-PA|Obp99b-RA	10.32	948	956
CG7593-PA|CG7593-RA	9.23	1189	1197
CG7598-PA|CG7598-RA	9.06	76	84
CG7643-PA|Ald-RA	9.67	1489	1497
CG7650-PA|CG7650-RA	10.32	874	882
CG7714-PA|CG7714-	11.21	1476	1484

RA			
CG7717-PA|Mekk1-RA	11.21	41	49
CG7717-PB|Mekk1-RB	11.21	41	49
CG7730-PC|CG7730-RC	9.23	289	297
CG7770-PA|CG7770-RA	10.94	277	285
CG7771-PA|sim-RA	9.23	1335	1343
CG7771-PB|sim-RB	9.23	1152	1160
CG7773-PA|fidipidine-RA	9.40	942	950
CG7789-PA|CG7789-RA	10.94	874	882
CG7800-PA|CG7800-RA	9.40	483	491
CG7811-PA|b-RA	10.04	490	498
CG7849-PA|CG7849-RA	9.40	592	600
CG7849-PB|CG7849-RB	9.40	592	600
CG7980-PA|RabX5-RA	9.67	1035	1043
CG7994-PA|CG7994-RA	9.67	804	812
CG8001-PA|CG8001-RA	10.32	522	530
CG8016-PA|rad201-RA	10.32	551	559
CG8055-PA|CG8055-RA	10.32	1396	1404
CG8057-PB|CG8057-RB	9.51	1480	1488
CG8070-PA|Mys45A-RA	9.51	547	555
CG8091-PA|Nc-RA	9.06	1470	1478
CG8102-PA|CG8102-RA	9.40	734	742
CG8102-PB|CG8102-RB	9.40	734	742
CG8105-PA|CG8105-RA	9.06	499	507
CG8107-PA|CalpB-RA	9.40	770	778
CG8145-PA|CG8145-RA	10.04	9	17
CG8151-PA|Tfb1-RA	9.40	1382	1390
CG8151-PB|Tfb1-RB	9.40	1382	1390
CG8151-PC|Tfb1-RC	9.40	1382	1390
CG8180-PA|CG8180-RA	10.04	815	823
CG8189-PB|ATPsyn-b-RB	9.06	1435	1443

CG8197-PA|CG8197-RA	9.40		360		368	
CG8201-PA|par-1-RA	9.96		28		36	
CG8201-PB|par-1-RB	9.96		28		36	
CG8201-PL|par-1-RL	9.96		28		36	
CG8201-PN|par-1-RN	9.96		28		36	
CG8204-PA|CG8204-RA	9.96		912		920	
CG8228-PA|CG8228-RA	9.51		561		569	
CG8232-PA|CG8232-RA	10.94		572		580	
CG8233-PC|CG8233-RC	9.67		1223		1231	
CG8241-PA|CG8241-RA	9.67		403		411	
CG8245-PA|CG8245-RA	11.21		1444		1452	
CG8298-PA|CG8298-RA	9.40		702		710	
CG8298-PB|CG8298-RB	9.40		492		500	
CG8361-PA|HLHm7-RA	9.96		1355		1363	
CG8439-PA|Cct5-RA	9.51		443		451	
CG8439-PB|Cct5-RB	9.51		253		261	
CG8457-PA|Cyp6t3-RA	10.04	9.40	104	1044	112	1052
CG8476-PA|CG8476-RA	11.21		796		804	
CG8525-PA|CG8525-RA	10.04		147		155	
CG8548-PA|Kap-alpha1-RA	9.96		956		964	
CG8568-PA|CG8568-RA	10.32		245		253	
CG8571-PA|smid-RA	9.06		745		753	
CG8571-PB|smid-RB	9.06		212		220	
CG8610-PA|Cdc27-RA	11.21		1232		1240	
CG8622-PA|Acp53Ea-RA	9.67		297		305	
CG8624-PA|melt-RA	9.40		111		119	
CG8624-PB|melt-RB	9.40		111		119	
CG8630-PA|CG8630-RA	9.96		1484		1492	
CG8641-PA|CG8641-RA	10.32		958		966	
CG8671-PA|CG8671-RA	11.21		824		832	
CG8671-PB|CG8671-RB	11.21		824		832	

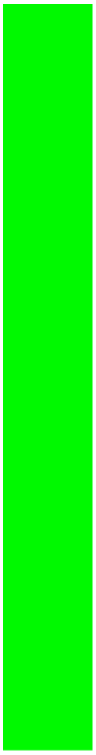
CG8674-PA|l(2)k14505-RA	9.51		620		628	
CG8694-PA|LvpD-RA	10.04	9.40	347	724	355	732
CG8696-PA|LvpH-RA	10.04	9.40	882	505	890	513
CG8765-PA|CG8765-RA	9.51		246		254	
CG8774-PA|CG8774-RA	9.40		455		463	
CG8774-PB|CG8774-RB	9.40		284		292	
CG8776-PA|CG8776-RA	11.21		516		524	
CG8793-PA|CG8793-RA	10.32		324		332	
CG8843-PA|sec5-RA	9.67		920		928	
CG8869-PA|Jon25Bii-RA	9.06		125		133	
CG8873-PA|CG8873-RA	9.96		1483		1491	
CG8874-PC|Fps85D-RC	9.67		845		853	
CG8877-PA|CG8877-RA	9.96	9.51	79	627	87	635
CG8905-PA|Sod2-RA	10.04		729		737	
CG8933-PA|exd-RA	9.51		1021		1029	
CG8933-PB|exd-RB	9.51		1021		1029	
CG8933-PC|exd-RC	9.51		1021		1029	
CG8937-PA|Hsc70-1-RA	9.40		1181		1189	
CG8942-PA|CG8942-RA	9.06		165		173	
CG8958-PA|CG8958-RA	9.67		1106		1114	
CG8998-PA|Roc2-RA	9.40		160		168	
CG9012-PA|Chc-RA	9.06		1168		1176	
CG9012-PB|Chc-RB	9.06		1168		1176	
CG9012-PC|Chc-RC	9.06		1168		1176	
CG9012-PD|Chc-RD	9.06		1168		1176	
CG9021-PA|CG9021-RA	9.40		835		843	
CG9072-PA|CG9072-RA	9.23		1124		1132	
CG9075-PB|eIF-4a-RB	9.06		849		857	
CG9075-PD|eIF-4a-RD	9.06		849		857	
CG9094-PA|CG9094-RA	9.23		1261		1269	
CG9094-PB|CG9094-RB	9.23		1261		1269	

CG9127-PA|ade2-RA	10.94		1303		1311	
CG9127-PB|ade2-RB	10.94		1303		1311	
CG9127-PC|ade2-RC	10.94		1303		1311	
CG9133-PB|CG9133-RB	9.51		1231		1239	
CG9133-PC|CG9133-RC	9.51		1231		1239	
CG9133-PD|CG9133-RD	9.51		1089		1097	
CG9168-PA|CG9168-RA	11.21		358		366	
CG9201-PB|Grip128-RB	9.96		431		439	
CG9203-PA|CG9203-RA	9.96	9.40	119	364	127	372
CG9211-PA|iHog-RA	9.51		642		650	
CG9214-PA|Tob-RA	9.67		791		799	
CG9214-PB|Tob-RB	9.67		791		799	
CG9246-PA|CG9246-RA	9.40		737		745	
CG9254-PA|CG9254-RA	9.51		1051		1059	
CG9263-PA|CG9263-RA	9.67		117		125	
CG9277-PA|betaTub56D-RA	10.04		1003		1011	
CG9277-PC|betaTub56D-RC	10.04		180		188	
CG9277-PD|betaTub56D-RD	10.04		180		188	
CG9300-PA|CG9300-RA	10.94		613		621	
CG9325-PA|hts-RA	9.67		747		755	
CG9325-PB|hts-RB	9.67		747		755	
CG9325-PC|hts-RC	9.67		747		755	
CG9325-PD|hts-RD	9.67		747		755	
CG9325-PE|hts-RE	9.67		747		755	
CG9325-PF|hts-RF	9.67		747		755	
CG9328-PA|CG9328-RA	9.96		1188		1196	
CG9342-PA|CG9342-RA	9.23		679		687	
CG9346-PA|CG9346-RA	10.32		787		795	
CG9379-PA|by-RA	9.40		309		317	
CG9392-PA|CG9392-RA	10.32		508		516	
CG9436-PA|CG9436-RA	9.06		1304		1312	

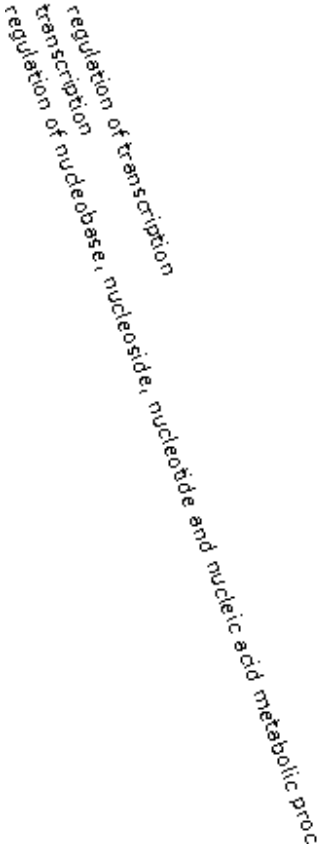
CG9441-PA|Pu-RA	9.40	191	199
CG9441-PC|Pu-RC	9.40	191	199
CG9444-PA|CG9444-RA	11.21	1393	1401
CG9445-PA|CG9445-RA	9.96	1140	1148
CG9454-PA|CG9454-RA	10.04	884	892
CG9455-PA|CG9455-RA	10.04	1062	1070
CG9458-PA|CG9458-RA	11.21	1100	1108
CG9483-PA|CG9483-RA	9.06	970	978
CG9543-PA|CG9543-RA	10.32	743	751
CG9571-PA|CG9571-RA	9.51	387	395
CG9576-PA|CG9576-RA	9.06	1483	1491
CG9587-PA|CG9587-RA	10.32	230	238
CG9610-PA|Poxm-RA	10.04	1189	1197
CG9610-PB|Poxm-RB	10.04	621	629
CG9653-PA|brk-RA	10.32	451	459
CG9660-PD|toc-RD	10.04	1294	1302
CG9672-PA|CG9672-RA	9.06	1345	1353
CG9682-PA|CG9682-RA	10.04	115	123
CG9699-PG|CG9699-RG	10.94	584	592
CG9707-PA|Acox57D-p-RA	9.23	70	78
CG9825-PA|CG9825-RA	10.94	1041	1049
CG9864-PA|CG9864-RA	9.23	71	79

APPENDIX 3.2

Annotation Cluster 2

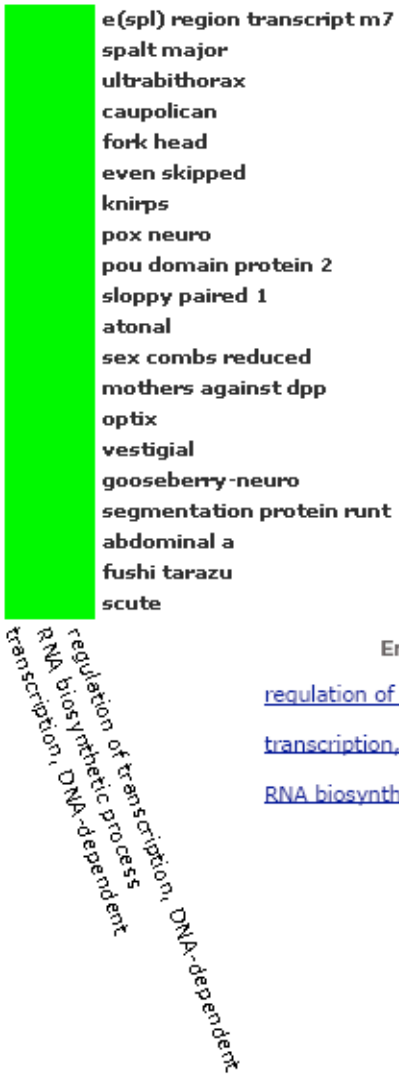


e(spl) region transcript m7
glial cells missing
scute
spalt major
ultrabithorax
caupolican
fork head
even skipped
bifid
knirps
pox neuro
pou domain protein 2
sloppy paired 1
atonal
rh50880p
sex combs reduced
mothers against dpp
optix
vestigial
anterior open
gooseberry-neuro
segmentation protein runt
abdominal a
fushi tarazu



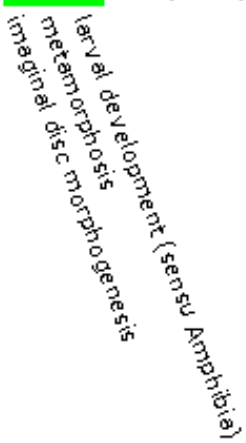
Enrichment Score: 8.83	Count	P_Value	Benjamini
regulation of transcription	24	3.9E-10	1.2E-7
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	24	2.4E-9	4.7E-7
transcription	24	3.5E-9	6.0E-7

Annotation Cluster 3



Enrichment Score: 6.75	Count	P_Value	Benjamini
regulation of transcription, DNA-dependent	20	5.8E-8	7.0E-6
transcription, DNA-dependent	20	3.0E-7	3.3E-5
RNA biosynthetic process	20	3.2E-7	3.3E-5

Annotation Cluster 4



Enrichment Score: 5.51	Count	P_Value	Benjamini
imaginal disc morphogenesis	13	2.1E-6	1.7E-4
metamorphosis	13	3.7E-6	2.6E-4
larval development (sensu Amphibia)	13	3.7E-6	2.6E-4

Annotation Cluster 5

Enrichment Score: 5.24	Count	P_Value	Benjamini
imaginal disc-derived appendage morphogenesis	10	5.1E-6	3.5E-4
imaginal disc-derived appendage development	10	5.8E-6	3.7E-4
appendage morphogenesis	10	5.8E-6	3.7E-4
appendage development	10	6.5E-6	4.1E-4

Chapter 4 – Investigation of the transcriptional regulation of *Six4* through the 3rd intron enhancer

4.1 Introduction

The two previous chapters have attempted to elucidate the role of *Six4* in *Drosophila* development through the identification of its downstream regulatory targets. Additionally, Chapter 1 offers a description of the *Six4* expression pattern which is then used in Chapter 3 to address the possibility of regulation of candidate genes by Six4. In a departure from the previous chapters the analyses presented herein will try to deconstruct the regulatory influences acting upon Six*4* itself in an attempt to place *Six4* in the tapestry of interactions that regulate various developmental events with particular emphasis on gonadogenesis (the formation of the male and female gonads). This chapter deals with the factors that provide the spatiotemporal information for *Six4* expression, as it is likely that their developmental function is mediated or refined through the intervention of Six4. The following section offers a description of the *Six4* expression pattern.

4.2 *Six4* expression

As stated previously, the *Six4* expression pattern was first described by Kirby et al. (2001). According to these authors, *Six4* is expressed during embryogenesis in the developing head region, the mesoderm and the CNS. Kirby et al. (2001) focus on the mesodermal expression of *Six4* because of the mesodermal heritage of the structures most affected in *Six4* null mutants ($Six4^{289}$/ $Six4^{289}$), the mesodermally derived musculature and the somatic gonadal precursors (SGPs, located in parasegments 10-12). Loss of Six4 also affects the fat body, although this issue is addressed in a later study by Clark et al. (2006). Kirby et al. (2001) describe *Six4* mesodermal expression as being segmental and then becoming confined to the SGPs until they associate with the Primordial Germ Cells (PGCs) at which point *Six4* expression in these cells is strengthened. *Six4* expression is reported to completely coincide with that of *eyes absent* (*eya*, see chapter 1) although expression of *eya* is broader. Fig. 4.1 is taken from Kirby et al. (2001) and summarises the authors' observations on *Six4* expression.

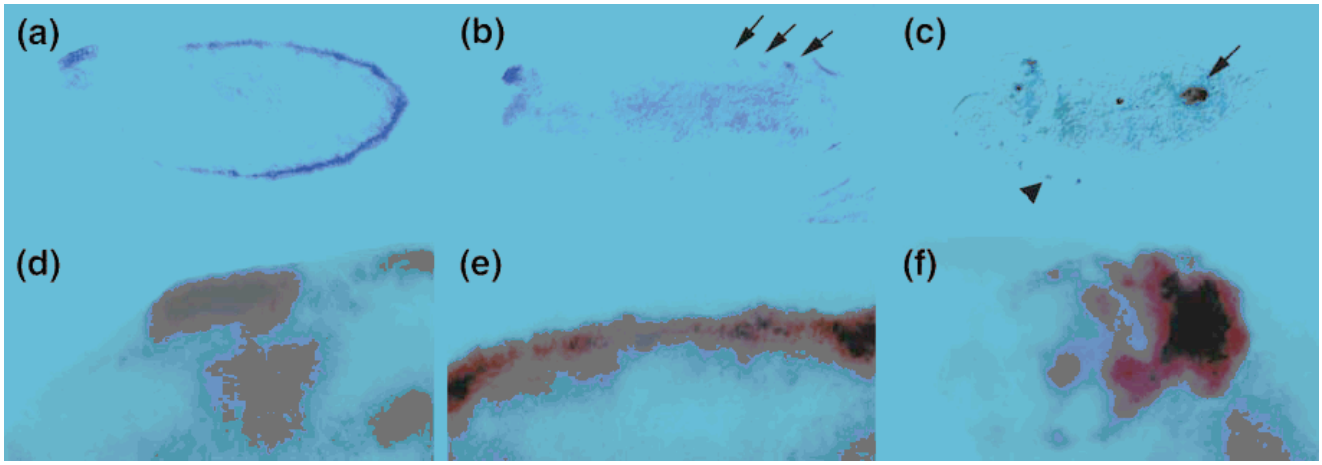


Fig. 4.1 A summary of the *Six4* expression pattern as described by Kirby et al. (2001) through in situ hybridization with an antisense *Six4* RNA probe. (a) at stage 9 *Six4* is expressed transiently throughout the mesoderm (a patch of expression in the developing head can also be seen). (b) at stage 13 this expression becomes confined to the SGPs (arrows) (some expression in the head is still visible). c) at stage 15, *Six4* expression in the now coalesced gonad continues (arrow). Some expression in the ventral nerve chord is also visible (arrowhead). (d-f) expression in the developing head (d), mesoderm (e) and the SGPs in the coalesced gonad (f) coincides with that of *eya*. Expression of *Six4* in these images is purple (seen here as black because of complete overlap with *eya* expression) whereas expression of *eya* is pink.

The *Six4* expression pattern has further been characterised by Clark et al. (2006). They report *Six4* expression in the trunk mesoderm from stage 9 but no expression in the cephalic mesoderm. Initial mesodermal *Six4* expression coincides with that of *Drosophila Mef2* (a major actor in mesodermal development and myoblast fusion, *Mef2* is not hit in the whole genome PWM scan described in chapter 3). By stage 10, *Six4* expression is limited to the ventral and lateral mesoderm until expression is only retained in the SGPs (stage 13).

Ivan Clark has attempted a deconstruction of the *Six4* expression pattern by using non-coding genomic regions in the vicinity of the *Six4* locus to drive GFP expression from reporter constructs. Based on his findings, he reports that mesodermal expression of *Six4* is solely driven by an enhancer located within the 3rd intron of the *Six4* gene. Similarly, the 4kb non-coding region directly upstream of the *Six4* transcriptional start can recreate the cephalic aspect of the *Six4* expression pattern (Ivan Clark, personal communication; Clark et al., 2006). The *Six4* 3rd intron expression pattern is summarised in Fig. 4.2 (taken from Clark et al., 2006). This study attempts to elucidate the transcriptional regulation of *Six4* through the 3rd intron enhancer (Six4-3int) using a combination of *in silico* and *in vivo* approaches. The following sections focus on the regulation of *Six4* through the 3rd intron enhancer.

Finally, Clark et al. (2006) propose a role for *Six4* in the development of the non-dorsal mesoderm that is analogous to that of *Six4* in the dorsal mesoderm (establishing dorsal mesoderm fates by specifying various dorsal mesodermal organ primordia).

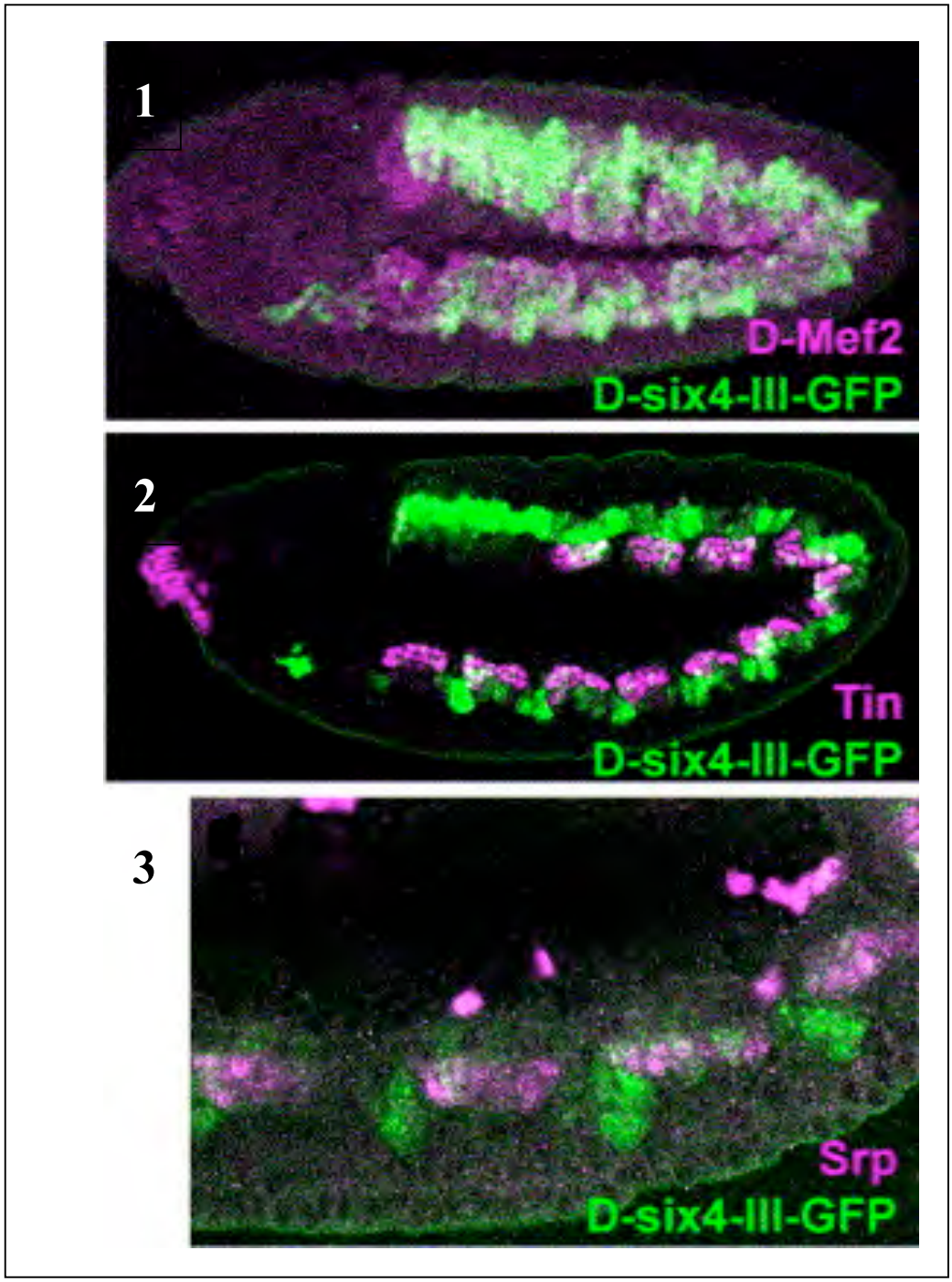


Fig. 4.2 Summary of the Six4-III-GFP expression pattern as described by Clark et al. (2006). **1)** Stage 10 mesodermal Six4-III-GFP expression overlaps that of D-Mef2 (magenta). **2)** At stage 10, Six4-III-GFP does not overlap with expression of dorsal Tin (magenta) supporting the argument of Clark et al. (2006) of Six4 being a mesodermal patterning factor with a role analogous and complementary to that of Tinman (see section 4.4). **3)** Stage 10 embryo at higher magnification stained for GFP and Srp (a dorso-lateral fat body cell marker). The involvement of Six4 in fat body development is inferred from the disruption of the fat body in Six4 null mutants. (Figure presented as published in Clark et al. 2006)

4.3 Six4-3int regulation analysis

Regulation of *Six4* mesodermal expression is controlled through TFBSs located within the Six4-3int enhancer. This section focuses on methods for identifying the genomic regions that may harbour such sites as well as the putative TFBSs within them.

As described previously, knowledge of the sequence of an enhancer element may provide information about its potential regulation through the identification of TFBSs using various *in silico* approaches (some such methods are also described and utilised in chapter 3). The theory underpinning most computational methods for the identification of cis-regulatory elements or modules (CREs or CRMs) is reviewed in Wasserman and Sandelin (2004). Most of the considerations and principles described therein still hold true today even if the used algorithms may have changed. Essentially, the regulatory potential of a suspected element is assayed based on a number of criteria. Firstly, putative TFBSs are identified through comparisons between genomic sequences and models (usually in the form of PWMs) that attempt to capture the sequence specific binding properties of transcription factors. PWM information is available in the form of databases such as TRANSFAC (900 PWMs, Matys et al., 2003) and JASPAR (138 PWMs, Vlieghe et al., 2006; Bryne et al., 2008), or occasionally through individual publications (for a few less commonly used databases see Wei and Yu, 2007). Most transcription factors have relatively short recognition sequences (usually 6-14 bp long) in the relative vicinity of the gene they regulate (Wasserman and Sandelin, 2004; Kadonaga, 2004). Given the short length of TFBSs and the, often degenerate, nature of many PWMs (which may reflect the relaxed binding preferences of their parent TFs) most PWM-utilising scanning algorithms generate large numbers of false positive hits. It is indeed conceivable that some of the false positive PWM hits may constitute genuine *in vitro* binding sites (*in vivo* binding may be heavily reliant on context, such as the presence of conformational factors that enable ligand-DNA binding). Various bioinformatics approaches have been utilised to counteract this problem and mine useful data out of non-coding sequences. Such approaches include analysis of gene expression profiles to determine the possibility of combinatorial gene regulation, and cis-regulatory module (CRM) detection methods as well as phylogenetic footprinting. Most of these methods have been used to some extent in chapter 3 to assess the validity of putative

Six4 regulatory targets. What follows is a focussed analysis of the Six4-3int enhancer using these methods.

It is worth mentioning that TFBS detection and/or Motif elicitation is not a simple process. The number of algorithms currently available that try (and occasionally succeed) to address these issues is indicative of the nature of this problem. Wei and Yu (2007) make mention of no less than 81 different TFBS and motif detection algorithms whereas Sandve and Dravlos (2006) report at least 100 (some 25 of which are used or mentioned in this study). Tompa et al. (2005) (see also the supplementary Li and Tompa, 2006) perform an evaluation of 13 motif detection algorithms and conclude that most algorithms are prone to generating many false positive results and that researchers should not rely on a single program and should investigate more than just the top scoring motifs. However, when combined with other approaches such as phylogenetic footprinting, TFBS predictions can provide useful information about gene regulation.

4.4 Six4-3int phylogenetic footprinting and shadowing analysis

This section addresses phylogenetic conservation of the Six4-3int enhancer and its potential significance towards the transcriptional regulation of Six4. The concept of fooprinting is visited in some detail in section 3.14 and is essentially based around the concept of evolutionary pressure acting upon functional TFBSs and conserving them, allowing for their detection through interspecific comparisons. The footprinted regions within the Six4-3int enhancer (Six4-3int) identified here are used in subsequent sections for detecting putative TFBSs.

The genomic sequence for Six4-3int was obtained through Flybase, version FB2008_05, released May 30, 2008 (Fig. 4.3). The interspecific conservation of the Six4-3int enhancer was initially assessed through a cursory comparison between the orthologous sequences in the MULTIZ multi-species genomic alignment available through the UCSC genome browser. Fig. 4.4 provides a visual representation of the extent of conservation of the Six4-3int sequence between the 12 *Drosophila* genomes, *A.mellifera*, *T.castaneum* and *A.gambiae*.

EVOprinterHD (Odenwald et al., 2005; Yavatkar et al., 2008) was used to generate an enhanced 3X-BLAT (BLAST-like alignment tool, Kent, 2002) alignment of the 12 *Drosophila* genomes and highlight footprinted and shadowed regions (regions that are either completely or partially conserved between the orthologous

sequences). The results of this analysis can be seen in Fig. 4.3. This analysis was performed to highlight areas of likely functionality within the Six4-3int enhancer and scan them for putative TFBSs.

The Six4-3int enhancer was found to contain a number of very highly conserved elements (footprints, see Fig. 4.3) in an otherwise relatively non-conserved sequence. 11 conserved sequence blocks (CSBs) were discovered ranging from 6-24bp (Table 4.1, the location of these CSB's with respect to the topology of the Six4-3int enhancer can be seen in Fig. 4.7). A CSB is a genomic region that is common to all the orthologous sequences that are used to generate a BLAT alignment. These CSBs where compared to other known *Drosophila* enhancer-type specific CSBs in order to identify their potential biological significance. What follows is a brief overview of the cis-Decoder utility that was used to analyse these CSBs and which is linked to EVOprinterHD.

gtgagttgaaatcatagattctcatgtcctcatccataaaatccactatataacttatccttttaactataaccc 75

acatcctatcttgaacacataaaatagaatttttattttaagaactgaaaatcgtcaaagaatcaatcaaaagatt 150

agttgtaattagagattattgactatttcctaaaaaacaactttgaaaagagctaataccatttgaagtgtagtt 225

agattctcttcacatttcttttgaacttgggccaaccaagttcgttaggaggggaatctagggcggcatggtaat 300

ttccac**TTCC**tC**ATAACTCGTAAA**ACAATTCATtcaaCaCTTGaGcgtgttgcgccccagaaccaaactccgaaa 375

tccgaaatgcgaacgaatatAATCA**ATT**g**TT**ctgg**AATTGGCATTAAAAAT**CGT**TT**g**TAAGTGACCGT**tccgtac 450

gg**CACTTG**tg**C**cTCcaggccgcc**CACTTGAGCGTGTTTTAT**G**TCCAATTT**cggccg**GC****TCCTT**gcacggtcCTcT 525

GCA**TATTT**A**TAGA**gcccagggactcgg**GAC****GTG**c**C**ccgactcctaacgcA**GCGCATGCGTGTAGCATGTGCC**cctc 600

atcctcctcaccattcccattctcctcctccgcCcgCatCCCGCTGtttgatcgcatcgcctctacagGGC**GTCA** 675

AATTGCAgcacataccacacccgagctcacaagtggagtaaatgtgttctttttagctgctcgaaaactcacagt 750

cgcacttccatgactagaaccatttggaaaacgcagagcgatcttaaaaaatggagtcatgcttcacttacattg 825

ataattgttaagggaattttaggggattccatagatagttccactgcctgaaactaaaA**CATATATTTATATAT** 900

AgtaactcTATAtTTTaagtgcTaTg**TTATAAA**TAT**A**T**TTAAAAA**Actaattgtttgattttag 975

Fig. 4.3 Primary sequence of the Six4-3int enhancer. Black upper case lettering represents sequences conserved between the 12 *Drosophila* species in the BLAT alignment. Colored bases represent sequences present in all species except *D.simulans*, *D.sechellia*, *D.yakuba*, *D.erecta*, *D.ananassae*, *D.pseudoobscura*, *D.persimilis*, *D.grimshawi*, *D.virilis*, *D.mojavensis* or *D.willistoni*

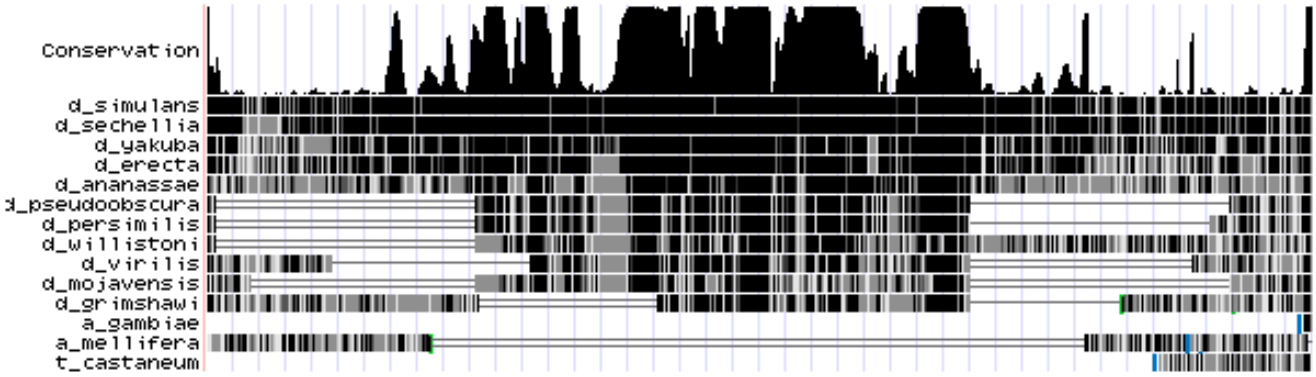


Fig. 4.4 Conservation chart of the MULTIZ 15 genome alignment of the Six4-3int enhancer available through the UCSC genome browser.

Brody et al. (2007) performed an analysis of 2,086 CSBs identified from 134 characterized enhancers (35 mammalian and 99 *Drosophila* enhancers, *Drosophila* enhancers obtained through REDFly2.0, see chapter 3) using EVOprinterHD. They maintain that CSBs may provide input for tissue-specific coordinate gene expression by harbouring TFBSs. These authors have identified numerous short (6-14 bp, consistent with most TFBS sizes) highly conserved DNA sequence elements, they called cis-Decoder tags (*c*DTs, see below for cis-Decoder), within these collected enhancer CSBs. These cDTs belong to two categories, i) those that are conserved only in enhancers known to confer a specific expression pattern (*Drosophila* libraries provided by these authors include mesodermal, segmental and neural enhancers) and ii) those that are common to divergently regulated enhancers. Brody et al. (2007) developed the cis-Decoder utility (http://evoprinter.ninds.nih.gov/cisdecoder/index.htm) to allow for comparisons to be made between these cDT libraries and other identified CSBs. ‘‘Because this approach does not rely on any previously described transcription factor consensus DNA-binding site information or any other predicted motif or the presence of overrepresented sequences, *cis*-Decoder analysis affords an unbiased 'evo-centric' view of shared single or multiple sequence homologies between different enhancers’’(Brody et al., 2007).

The 11 Six4-3int enhancer CSBs identified in this study were scanned for matches to cDTs conserved in other *Drosophila* mesodermal enhancers using the cDT-scanner application in the cis-Decoder program suite (25 enhancers including early and late embryonic), as well as those common to numerous enhancer types but found to be over-represented in mesodermal enhancers. The supplemental data in Brody et al. (2007) provide a complete list of the enhancers used to construct this list. Matches to 9 previously identified cDTs were reported within Six4-3int. These cDTs, along with hits to PWMs of known *Drosophila* TFs (see section 4.4) located within the scanned CSBs can be found on table 4.1 and Fig. 4.5.1 respectively.

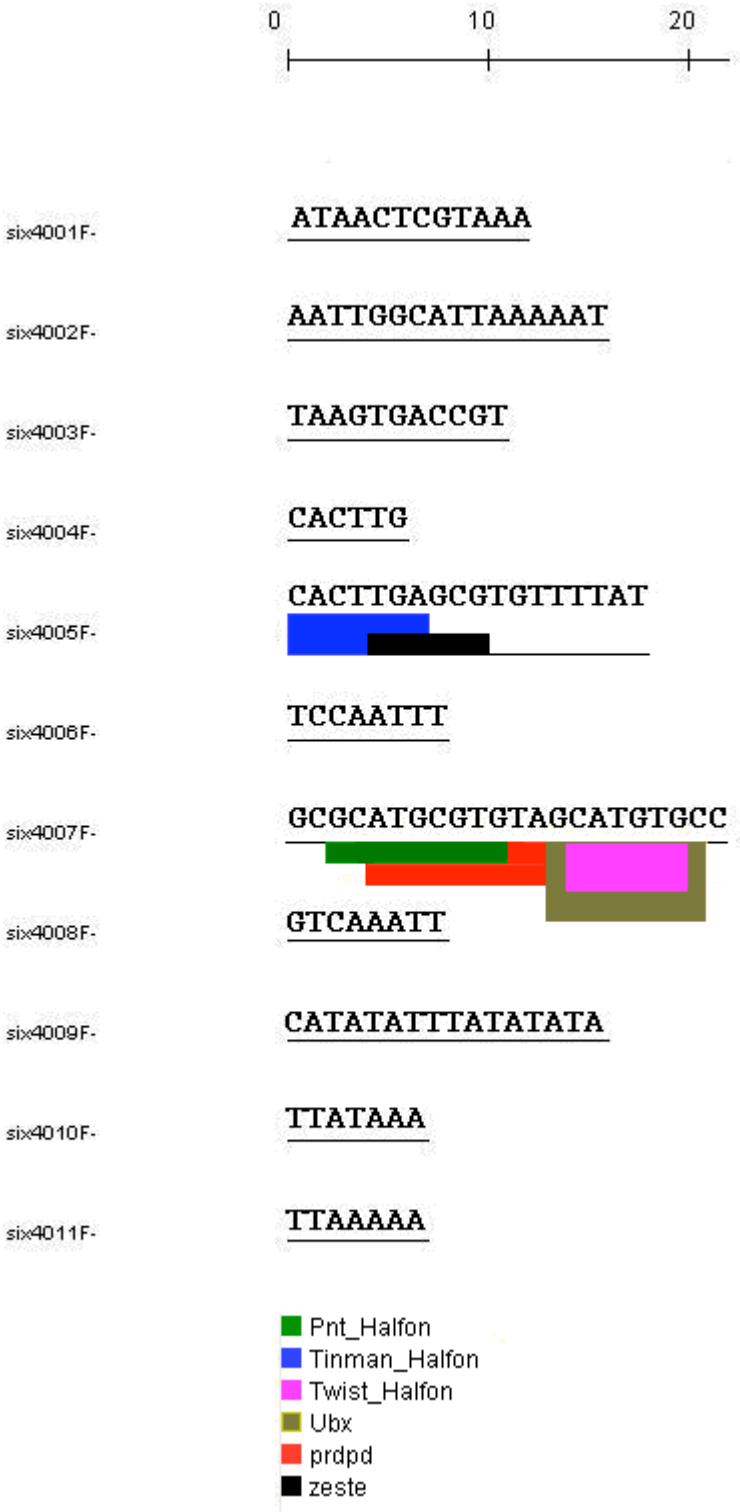


Fig. 4.5.1 The 11 CSBs discovered in the Six4-3int using an EVOprinter generated BLAT alignment. Hits to *Drosophila* TF PWMs within these CSBs are provided (as well as the publication of origin where relevant). Matrix scores generated by MotifScanner were: Pnt (48.59), Tinman (4733), Twist (211.95), Zeste (214), Prdpg (763.49), Ubx (385.39)

This analysis was performed because the Six4-3int enhancer is known to regulate mesodermal expression of *Six4* and could conceivably do so through the action of mesodermal specific TFs which could bind to the conserved sequences. Matches to 4 mesodermal specific and 5 mesodermally over-represented cDTs were detected. Table 4.1 summarizes these findings and presents the other enhancers that also share these conserved elements (and are therefore potentially co-regulated).

Some of the potentially co-regulated factors identified in this analysis have previously been associated with *Six4*. These factors are Tinman (Tin), Bagpipe (Bap) and Decapentaplegic (Dpp). Specifically, Clark et al. (2006) ‘‘propose that earlier in development (at stages 8/9), part of *tin*'s function ventrally is to initiate expression of *Six4’’*. This hypothesis is based on the observation that expression of a GFP reporter construct driven by the Six4-3int enhancer is severely reduced in *tin* mutant embryos. *Tinman* is involved in general mesodermal patterning (see also section 4.5.2). Recent unpublished findings by members of Eileen Furlong’s research group have established the direct regulation of Six4 by Tinman (Eileen Furlong, personal communication, I became aware of this after the conclusion of this study). Additionally, Decapentaplegic (Dpp) maintains *tin* expression in the dorsal mesoderm (Frasch, 1995). Dpp is responsible for visceral mesoderm specification (along with Wingless, Lee and Frasch, 2005). Finally, Bagpipe (Bap, a marker for visceral musculature precursors) is known to mediate the initial dorso-ventral positioning of SGPs through establishing the dorsal fate by downregulating *eya* (Boyle et al., 1997). As such, Bagpipe also plays a role in the specification of visceral mesoderm (Azpiazu and Frasch, 1993; Lee and Frasch, 2005). Boyle et al. (1997) also present a role for Bap as a repressor of SGP fate through repression of *eya* (a co-factor of Six4). They report that a higher number of *eya* expressing SGP cells can be observed in Bap mutants. The *bap* enhancer contains 3 of the 4 mesodermal specific cDTs also found in the Six4-3int, as well as 2 of the mesodermally enriched ones suggesting that co-regulation with *Six4* is likely. Finally, Serpent (Srp) is a fat body marker. The fat body is another structure affected in *Six4* mutants (Clark et al., 2006) and is derived from the same subset of cells that give rise to the SGPs.

The interplay between these and other factors is revisited in section 4.6. The links between the roles of these genes in mesodermal development, suggest that they may be subject to co-regulation, possibly through the conserved elements present in their enhancers. These CSBs will be revisited in the following section that addresses the issue of putative TFBS identification in the Six4-3int enhancer.

Conserved Sequence Block	Sequence	Matches to previously reported cDTs	Matches to enhancers of other mesodermal factors
six4001F	ATAACTCGTAAA	ACTCGT(1)(n01;s01;m03)	Srp, Scr, Bagpipe
six4002F	AATTGGCATTAAAAAT	CATTAAAA (2) (m6;n0;s0)	Scr(2,3,5,6,7),
		CATTAAAAA(3) (m4;n0;s0)	Bagpipe(2,3,6),
		ATTAAAAAT(4) (m3;n0;s0)	Toll-6(4,5,6,7),
		GCATTAA (5)(n01;s00;m04)	Dpp(2,5),Pdp1(4),
		ATTAAAAA (6)(n01;s01;m05)	Tinman(7)
		TTAAAAAT (7)(n01;s01;m04)	
six4003F	TAAGTGACCGT	TGACCG (8) (m2;n0;s0)	Pdp1, Toll-6
six4004F	CACTTG		
six4005F	CACTTGAGCGTGTTTTAT		
six4006F	TCCAATTT		
six4007F	GCGCATGCGTGTAGCATGTGCC		
six4008F	GTCAAATT		
six4009F	CATATATTTATATATA	ATTTATA (9)(n01;s00;m04)	Scr, Toll-6, Bagpipe
six4010F	TTATAAA		
six4011F	TTAAAAA		

Table 4.1 Table of the 11 CSBs identified in the Six4-3int enhancer. Sequences and matches to previously reported cDTs are provided. cDT coding denotes their occurrence in the 3 different *Drosophila* enhancer categories. For example the *c*DT ACTCTG (1) is annotated n01;s01;m03-ACTCTG to indicate that there was 1 hit on a *Drosophila* neural CSB library, 1 hit on a segmental CSB library and 3 hits on a mesodermal CSB library. The genes linked to the enhancers containing CSBs that match these cDTs are also provided.

4.5 TFBS and putative regulatory element identification

The identified CSBs have been used to identify potential TFBSs. A number of algorithms have been developed for the prediction of TFBSs using PWMs (as well as other methods such as HMMs, see MAPPER, Marinescu et al., 2005). Das and Dai (2007) and Wei and Yu (2007) survey a number of motif finding utilities. Simple, PWM-utilising, algorithms, as well the considerations involved in their use are described in chapter 3. Most authors conclude that TFBS predictions are largely uninformative in the absence of additional data. Klepper et al.'s (2008) assessment of 8 composite motif discovery methods, although inconclusive, highlights the fact that most available algorithms are easily confused by the ''noise'' present in most datasets. In light of this, most researchers opt for the use of a combinatorial approach that utilises phylogenetic footprinting (addressed in the previous section) and enhancer analyses of co-expressed (and therefore potentially co-regulated) genes using PWM scanning algorithms. Section 4.6 presents the compilation of a list of co-expressed genes that will subsequently be used in such an analysis.

Most of the analyses described herein were performed using the TOUCAN 2 (version 3.1.0) regulatory sequence analysis platform (Aerts et al., 2005) except when operational restrictions (insufficient memory) necessitated the independent use of the individual algorithms used by TOUCAN 2.

4.5.1 Unfiltered MotifScanner analysis of Six4-3int

An initial analysis of the Six4-3int enhancer using the MotifScanner utility (version 3.1.1, incorporated into the TOUCAN 2 program suite) returned a large number of hits. Fig. 4.5.2 presents the results of this analysis. MotifScanner is based on the MotifSampler algorithm which is ''a Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes'' (Thijs et al. 2002). The PWMs used in this analysis include all the *Drosophila* matrices included in TRANSFAC and JASPAR as well as those included in Lifanov et al. (2003) (see also Papatsenko et al., 2002), Halfon et al. (2008) and Rajewsky et al. (2002). Unlike the analysis presented in chapter 3 where an informed decision can be made when determining the classifier threshold of a PWM, sampling algorithms that utilise large PWM libraries cannot set their cut-off points on the basis of individual scores (since matrices have different information contents and generate different scores) but do so on the basis of E-values (or other variables dependent thereupon). MotifScanner can

use higher-order Markov processes to model the non-coding sequence background and inform its expectation values. All analyses using MotifScanner described in this study use a pre-compiled *Drosophila* model of the 1st markov order (*Drosophila* EPD 1, essentially a representation of the average promoter composition, see chapter 3 for an explanation of markov orders) calculated from *Drosophila* promoters stored on the Eukaryotic Promoter Database (EDT Release 94, http://www.epd.isb-sib.ch/, Schmid et al., 2006). Sampling stringency is determined by a ‘‘prior’’ value that determines the number of hits one expects to find in the given search space. Higher prior values have lower stringency and generate more (false positive) hits. The value of this parameter depends on the size of the sequence being sampled. In this study all analyses of sequences ≤ 1kb used a prior value of 0.4 whereas all sequences ≤ 0.1kb used a prior value of 0.1 as per the suggestion of Herbert Mayer (TOUCAN 2 online tutorial). All other sequences use a prior value of 0.5. An analysies of the entire Six4-3int sequence using the parameters described here returned hits to 27 known TFs. Fig. 4.5.2 summarises all the reported putative TFBSs. Given the previous knowledge on *Six4* regulation, only a few of these hits are likely to correspond to regulators of *Six4*. It is therefore necessary to discern which of these putative TFBSs are likely to be responsible for Six4 regulation. The following sections describe ways of reducing the noise generated by TFBS analyses.

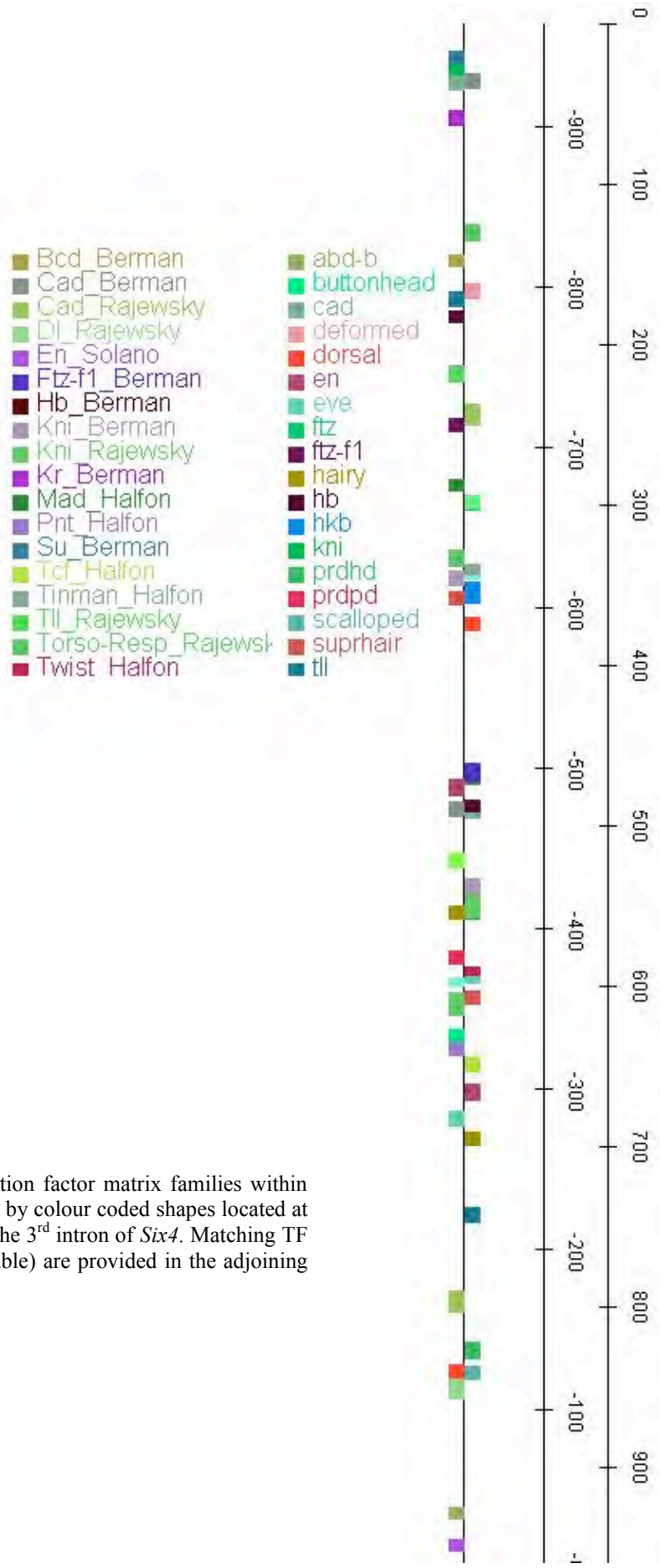


Fig 4.5.2 Map of matches to known transcription factor matrix families within the Six4-3int enhancer. Matches are designated by colour coded shapes located at their relative position (and orientation) within the 3rd intron of *Six4*. Matching TF PWMs and publication of origin (where available) are provided in the adjoining map key.

4.5.2 Identification of Footprinted putative TFBSs

In this section only the putative TFBSs reported in the 11 identified CSBs were considered for a TFBS analysis. 6 hits to 6 different TF PWMs were reported using MotifScanner. These hits are presented in Fig. 4.5.1 Putative TFBSs included those of Tinman (Tin, a known regulator of *Six4*), Twist (Twi, a predominantly mesodermal transcriptional activator that is required for gastrulation and mesoderm formation and confers expression consistent with the Six4 pattern, Thisse et al., 1987) Ultrabithorax (Ubx, another transcription factor involved in mesodermal cell fate specification that has an expression pattern also consistent with early *Six4* expression, Ponzielli et al., 2002) and Pointed (Pnt, another TF expressed in the mesoderm, Scholz et al., 1993). In contrast the putative TFBSs of the TFs *Paired* (Prd-paired domain, limited mesodermal expression, Gutjahr et al., 1994; Xue and Noll, 2002) and Zeste (very limited embryonic expression, Pirrotta et al., 1988; Benson and Pirrotta, 1988) are incompatible with *Six4* expression and therefore unlikely to be it's regulators.

It is, however, likely that some of the other reported TFBSs that were located outside the footprinted regions are of biological significance. The following sections address this possibility and try to use the knowledge of other genes that either interact with *Six4* or share a common expression pattern to make inferences on the possible functionality of these putative TFBSs.

4.6 Candidates for *Six4* co-regulation

This section deals with the compilation of a list of genes that share part of the expression pattern of *Six4* in an attempt to identify the regulatory cues that control *Six4* mesodermal expression. As described in chapter 1 as well as in section 4.2, *Six4* is expressed in a wide variety of cell types during embryo development. In this study I have chosen to focus on the expression of *Six4* in the mesoderm and more specifically on the somatic gonadal precursors (SGPs). This is an informed decision that was based on the fact that a search for all the genes that are co-expressed with *Six4* at any stage of development would likely yield a large number of results that could confuse any subsequent search for specific TFBSs. In contrast the SGPs are a small subset of cells that express a more manageable collection of genes and are as such more likely to provide insight on the regulation of *Six4*.

The *Drosophila* gonad is formed from the union of cells that derive from different cell lineages. The initially segregated primordial germ cells (PGCs, known as pole

cells in the earliest developmental stages) migrate into the embryo and contact the SGPs (the somatic component of the gonad) and coalesce into the immature gonad. The colonisation of the right location by PGCs is ensured through programmed cell death (mediated by the *Drosophila* p53 in association with the *outsiders* gene) of any mis-localising cells (Yamada et al., 2008).

For a detailed overview of the developmental events that characterise this process as well as some of the factors involved see Starz-Gaiano and Lehmann (2001), Santos and Lehnmann (2004) and Clark et al. (2007). A brief overview of this process is provided here for convenience. Initially the PGCs are the first cells to be specified in the posterior pole of the *Drosophila* embryo from maternally provided cytoplasm close to the anlage of the posterior midgut. They are under the control of a genetic pathway that is distinct from the one that controls somatic cell formation. They are then passively carried along the gut cavity by the extension of the germ band during the process of gastrulation. They are then carried into the embryo through the invagination of the posterior midgut primordium (PMP). They then actively migrate through the surface of the PMP and contact the SGPs in the gonadal mesoderm. These first steps are thought to be independent from Six4 involvement. Six4's role in gonadogenesis involves the specification and maintenance of the SGPs. During germ band retraction the PGCs (now paired to the SGPs) travel anteriorly until they colonise parasegments 10-12. Then at stage 15 both cell types coalesce to form the embryonic gonad.

4.6.1 Literature derived co-regulation candidates

A number of genes are known to be involved in the process of gonadogenesis. These include, from a *Six4* point of view,: *eyes absent* (*eya*, encoding a SIX family co-factor, see chapter 1), a gene required for SGP fate maintenance (Boyle et al., 1997; Broihier et al., 1998) along with *tinman* (*tin*, a potential regulator of Six4 according to Clark et al., 2006; also see Moore et al., 1998_2; Bodmer, 1993; Boyle and Dinardo, 1995), *abdominal a* and *b* (*abd-A* and *abd-B*, factors that determine competence of a parasegment for SGP induction and define the clusters of the lateral mesoderm that give rise to the gonadal mesoderm, Boyle and Dinardo, 1995; Cumberledge et al., 1992), *HMGCoA reductase* (*Hmgcr* or *Columbus*, a gene responsible for guiding the PGCs to the SGPs, Santos and Lehmann, 2004) and *zfh-1* (another mesodermal attractant for the PGCs and a suspected regulator of Six4

according to Clark et al., 2006). Given the involvement of these genes in the development of the gonadal mesoderm it was considered that at least some of these factors are subject to co-regulation. The regulatory regions of these genes (where available through REDFly2.0) were subjected to a TFBSs analysis to account for the presence of over-represented putative TFBSs. This gene enhancer collection was designated LDL (literature derived library). Details about these genes and others as well as their involvement in gonadogenesis can be found in Santos and Lehmann (2004).

Out of these genes, *eya*, *Hmgcr* and *Six4* have the most similar expression patterns. Eya, consistent with its role as a co-factor of *Six4*, has an expression pattern that echoes the mesodermal expression of *Six4* almost completely. It is very likely that expression of these two genes is controlled by the same factors. Similarly Clark et al. (2007) suggest a role for *Six4* as a regulator of *Hmgcr* (no *Hmgcr* expression is detected in $Six4^{289}$ homozygotes). Interestingly, no hits to the Six4 PWM were scored in the putative regulatory regions of these 2 genes (potentially hinting towards indirect regulation, or regulation through a TFBS lying outside the scanned genomic areas). Therefore, based solely on the PWM scan of the putative regulatory regions of *eya* (1 kb upstream of transcriptional start as well as the intronic regions) using the parameters outlined in chapter 3 no evidence of a putative TFBS for Six4 has been found near the *eya* locus. The possible reasons for these include the inability of the Six4 PWM to detect a Six4 TFBS, the potential existence of such a TFBS outside the searched genomic areas, or the regulation of *eya* through the mediation of another factor. Given the close relationship between these 3 genes their suspected enhancer elements were also subjected to a separate analysis (see below).

Additionally, a phylogenetic footprinting analysis of the sequences 1kb upstream of the *Hmgcr* and *eya* loci revealed 25 and 28 CSBs respectively for each of these genes (EVoprinter, parameters were set as described in the previous sections). This was done because no CRM is reported for *Hmgcr* and the enhancer for *eya* reported by Bui et al. (2000) controls *eya* expression in the eye and was not used in favour of the *eya* upstream region. An alignment between those CSBs and the ones detected in the Six4-3int enhancer (using the cis-Aligner utility in the cis-Decoder program package) showed few matches between the Six4-3int enhancer and the other 2 upstream regions. No hits to the PWMs of the TFs found in the CSBs of the Six4-3int enhancer were found in the CSBs of the upstream regions of the other 2 genes. Given

the existing knowledge on the nature of footprinting this finding does not constitute proof of the absence of co-regulation of these 3 genes.

4.6.2 Co-regulation candidates based on expression ontology

Finally, all genes reported to be expressed in the fat body/gonad primordium in the BDGP in situ database (Release 2, March 2007, Tomancak et al., 2002) were also subjected to an enhancer analysis in a search over-represented putative TFBSs. This was done to highlight common regulatory elements in the enhancers of genes that share parts of the Six4-3int expression pattern. I initially intended to conduct an enhancer analysis of all the genes expressed in the gonadal mesoderm in an attempt to identify regulating factors that may control *Six4* expression in the SGPs. To that end I scanned the in situ database at BDGP (Tomancak et al., 2002) for related ImaGO terms (the in situ project's equivalent of a GO entry, indicates expression in the corresponding part of the embryo) and genes associated with them. An ImaGO term does exist for the gonadal mesoderm (FBbt:00000135) but as of yet no genes are associated with it. The closest term is the fat body/gonad primordium (the part of the mesoderm that gives rise to both the gonadal mesoderm and the fat body, FBbt:00005520). 56 genes are currently associated with this term (Table 4.2). Enhancer elements for these genes were obtained through the REDFly2.0 database. For all genes with no REDFly2.0 entries the region 1kb upstream from the start codon was obtained through the automated sequence retrieval function of TOUCAN 2. Availability (on public-access databases) of verified enhancer elements is reported in Table 4.2. This list was also supplemented with the genes included in the LDL library. This is because the collection of genes with associated ImaGO data is not all-inclusive and only accounts for genes for which in situ hybridisation data is available. Members of the LDL collection are known to be expressed in the fat body/gonad primordium and can therefore be included in this list. This enhancer library was designated FGPL (fat body/gonad primordium library).

Additionally, a manual screening of images for the genes represented in FGPL revealed some candidates that mimicked the expression pattern of the Six4-3int enhancer more closely than the other entries. These genes and transposable element insertions have expression patterns that are almost indistinguishable from that described for Six4-3int by Clark et al. (2006). One such case (that of transposon

Stalker4{}1446) is showcased in Fig. 4.6. This enhancer library was designated CCEL (closely co-expressed library) and is described in table 4.3.

It is worth mentioning that 2 transposable elements with expression patterns matching that of *Six4* are included in these collections. One of these is the *412* transposable element (Costas et al., 2001) which has long been known to be expressed in the SGPs and is often used as an SGP marker (Kirby et al., 2001). The other is the *Stalker4* retrotransposon. Little is known about this transposable element but as can be seen in Fig. 4.6, its expression in the mesoderm is extremely similar to that of *Six4*. It has been shown that the expression of transposable elements is often independent of genomic context and must therefore be controlled by CRMs within the sequence of the transposon itself. Because of this, only the genomic sequence of these transposons was included in the enhancer libraries that contain them. No hits to the Six4 PWM have been found in either of these sequences hinting towards co-regulation with *Six4* rather than direct regulation by it.

A final enhancer library was assembled using the enhancers of the genes found to be associated with the ImaGO term ‘‘gonadal sheath’’ (FBbt:00004859). This term had 13 genes (as well the *Stalker4* transposon) associated with it (Table 4.4). Given the small number of genes expressed in the gonadal sheath this list was deemed to be more likely to contain genes subject to co-regulation. This list was supplemented with the Eya putative enhancer and Six4-3int and was designated GSL (gonadal sheath library).

4.7 TFBS analysis of compiled enhancer libraries

All the enhancer libraries described above were subjected to a TFBS analysis using MotifScanner (parameters as above) and the observed TFBS frequencies were compared to those obtained for the control enhancer set (all *Drosophila* entries in the Eukaryotic Promoter Database, -499 to +100 around TSS, scanned using MotifScanner, 0th order background model, prior 0.2). Since most of the sequences scanned were 1kb long (due to the all too frequent lack of availability of experimentally validated CRMs related to the genes of interest), the use of a prior of 0.2 when calculating the background frequencies of putative TFBSs (from 0.6 kb-long sequences) was deemed acceptable when compared to the value of 0.4 which was used when scanning the (mostly) 1kb-long sequences. Additionally, the use of a lower prior value when establishing background frequencies can only overstate the

statistical significance of any ove-representation that is detected. However, as seen below, most results were found to be statistically insignificant despite this fact. The results for these analyses can be found on tables 4.5.*.

Gene/element symbol	Biological Function	Availability of experimental enhancer data and publication of origin
CG12094		No
BcDNA:GH02419		No
CG30115		No
odd		Yes (Berman et al., 2002)
fat-spondin		No
scf		No
CG8745		No
beat-IIIc		No
CG8036		No
CG31361		No
CG6870		No
mael		No
Smn		No
CG9837		No
smid		No
CG14693		No
TepIV	Protease inhibitor activity	No
CG3011		No
CG10924		No
FK506-bp1	Protein folding	No
Stalker4	Transposable element	
Ahcy89E		No
CG8791		No
CG8286		No
fus		No
ppl		No
CG3999		No

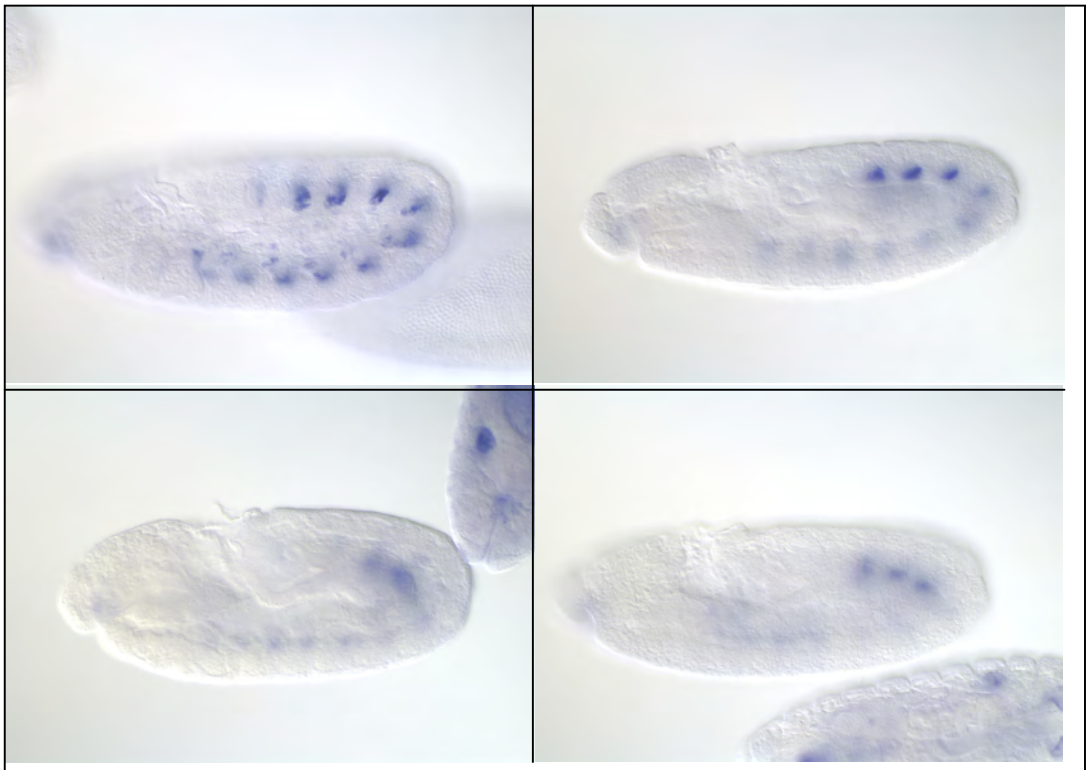
CG31160		No
ade2		No
CG9057		No
dsx		No
CG15828		No
CG13845		No
CG6910		No
Jheh1	Epoxide hydrolase activity	No
CG6415		No
CG11337		No
CG5241		No
412	Transposable element	No
Tapdelta	Protein retention in ER	No
ERp60		No
CG8147		No
ps	mRNA binding (spliceosome)	No
CG6934		No
eEF1delta	Translational elongation	No
bgm		No
pont		No
srp		Yes (Miller et al., 2002)
ade5		No
CG2003		No
CG6393		No
CG11315		No
La		No
abd-A	Specification of gonadal mesoderm (Boyle and DiNardo, 1995)	Yes (Shimell et al., 2000)
eya	Required for SGP fate maintenance, *Six4* co-factor (Boyle et al., 1997; Broihier et al., 1998)	No (Bui et al., 2000 report an eye enhancer that is of little interest to this study)

bap		Yes (Lee and Frasch, 2005)
tin	Required for SGP fate maintenance	Yes (Venkatesh et al., 2000)
dpp		Yes (Blackman et al., 1991)
Yp1	Fat body marker	Yes (Burtis et al., 1991; Garabedian et al., 1986)
zfh-1		No
Hmgcr		No
abd-B	Specification of gonadal mesoderm (Boyle and DiNardo, 1995)	Yes (Busturia and Bienz, 1993)

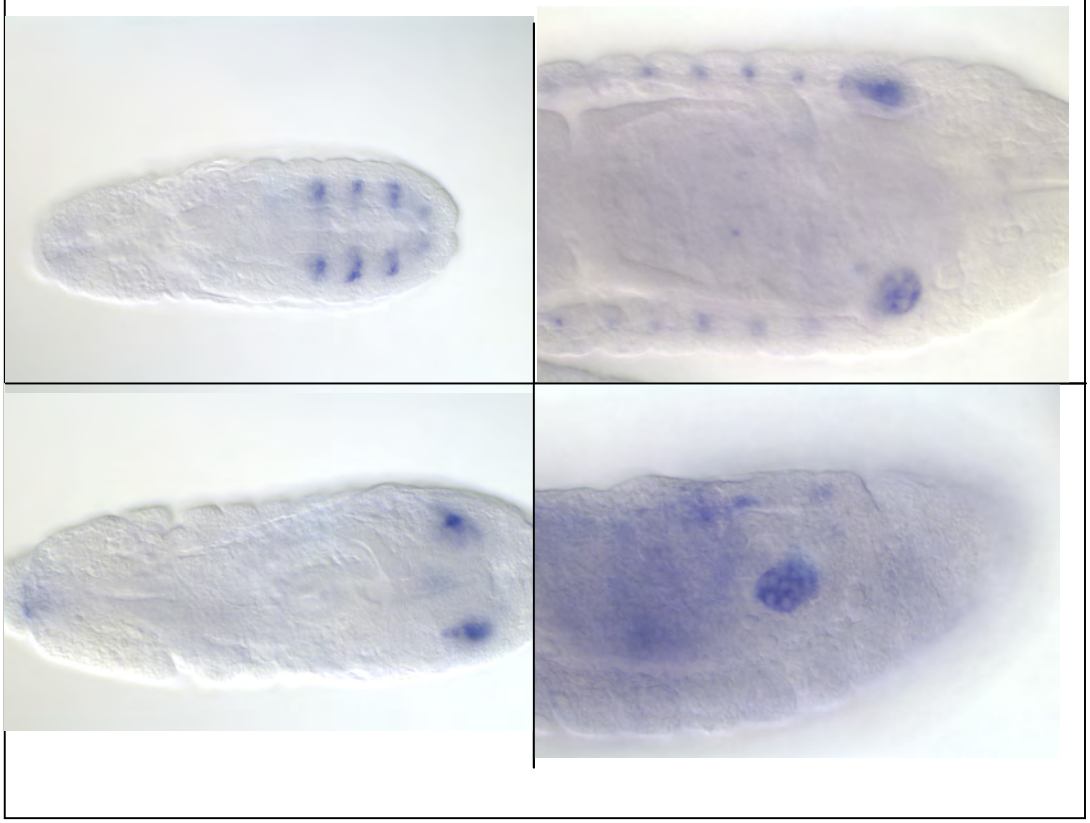
Table 4.2 FGPL (fat body/gonad primordium library). Proteins associated with the fat body/gonad primordium ImaGO term (FBbt:00005520). Biological functions for these proteins are included when available and potentially relevant to the role of Six4 in development.

Gene Name	**Biological function**
Stalker4{}1446	transposable element
412 retrotrasposon (412{}bw^{1})	transposable element
eya	Six4 co-factor
Hmgcr	gonad development and cholesterol metabolism
Gapdh1	Glycolysis
CG10082	phosphate metabolism and transport
CG6934	Unknown
Dpr17	Unknown

Table 4.3 CCEL (closely co-expressed library). Proteins found to be expressed in a pattern similar to that of Six4-3int driven GFP expression. Biological functions for these proteins are included when available.



Developmental stages: **7-9**(top left), **9-11**(top right), **11-13** (bottom left-right)



Developmental stages: **12-13** (top left), **13-16** (top right, bottom left-right)

Fig. 4.6 In situ hybridisation of *Stalker4* RNA (as found in the BDGP Gene Expression database). The pattern of expression closely mimics that conferred by the Six4-3int enhancer. The expression patterns of the other selected genes closely resemble the one seen above.

Gene Name	Biological function
CG12819	
CG11459	
l(2)08717 (CG15095)	
CG6934	
CG3074	
CG9989	
CG18279	
Stalker4{}1446	Transposable element
CG6225	
CG14693	
CG6188	
CG31361	
CG12094	
eya	Six4 co-factor
Six4	Transcription Factor

Table 4.4 GSL (Gonadal Sheath list). List of genes associated with the Gonadal Sheath ImaGO term (FBbt:00004859). Biological functions are included where available (and relevant).

Feature Name	n	Prob(occ{b}>=n	SIG
pho	36	0.033	-0.24
mad	45	0.241	-1.106
en	43	0.345	-1.262
cad	39	0.397	-1.323
ttk	27	0.536	-1.454
Pnt_Halfon	59	0.561	-1.473
Kr_Rajewsky	40	0.573	-1.482
hkb	25	0.683	-1.559
Torso-Resp_Rajewsky	23	0.767	-1.609
ftz-f1	55	0.780	-1.616
Ftz-f1_Berman	63	0.805	-1.63
Kr_Berman	29	0.853	-1.655
gt	16	0.886	-1.671
Mad_Halfon	77	0.891	-1.674
hairy	39	0.891	-1.674
Su_Berman	22	0.930	-1.693
kni	53	0.954	-1.704
prdpd	34	0.955	-1.704
Ftz_Berman	26	0.957	-1.705
buttonhead	76	0.975	-1.713

Table 4.5.1 CCEL. Closely co-expressed list

Feature Name	n	Prob(occ{b}>=n	SIG
gt	26	0.007	0.436
kni	57	0.100	-0.699
pho	23	0.167	-0.922
Bcd_Rajewsky	32	0.171	-0.933
Kr_Berman	30	0.181	-0.956
Kni_Berman	58	0.290	-1.162
ubx	43	0.296	-1.17
Pnt_Halfon	46	0.349	-1.242
tll	75	0.366	-1.262
Tcf_Halfon	56	0.387	-1.287
suprhair	84	0.388	-1.288
hb	47	0.419	-1.322
hairy	35	0.436	-1.338
Hb_Berman	50	0.439	-1.341
ttk	20	0.475	-1.376
zeste	78	0.480	-1.381
abd-b	72	0.503	-1.4
prdpd	32	0.527	-1.421
Tll_Rajewsky	50	0.562	-1.449
cad	26	0.567	-1.453

Table 4.5.2 FGPL. Fat Body/Gonad Primordium list

Feature Name	n	Prob(occ{b}>=n	SIG
gt	11	0.148	-0.91
cad	13	0.596	-1.515
pho	8	0.690	-1.579
ttk	8	0.749	-1.615
deformed	5	0.797	-1.642
Torso-Resp_Rajewsky	7	0.821	-1.655
en	11	0.860	-1.675
Kr_Berman	9	0.876	-1.683
Bcd_Rajewsky	9	0.915	-1.702
Dl_Rajewsky	11	0.932	-1.71
bcd	7	0.932	-1.71
Su_Berman	6	0.945	-1.716
ubx	13	0.962	-1.723
kni	16	0.971	-1.728
mad	8	0.975	-1.729
scalloped	15	0.981	-1.732
Bcd_Berman	12	0.984	-1.733
ftz-f1	13	0.987	-1.735
Cad_Rajewsky	12	0.991	-1.736
prdpd	8	0.991	-1.736

Table 4.5.3 GSL. Gonadal Sheath list.

Feature Name	n	Prob(occ{b}>=n	SIG
gt	11	0.060	-0.521
cad	13	0.353	-1.288
pho	8	0.494	-1.434
ttk	8	0.560	-1.489
deformed	5	0.653	-1.555
Torso-Resp_Rajewsky	7	0.661	-1.561
en	11	0.678	-1.571
Kr_Berman	9	0.720	-1.597
Bcd_Rajewsky	9	0.785	-1.635
Dl_Rajewsky	11	0.801	-1.644
bcd	7	0.829	-1.659
ubx	13	0.857	-1.673
Su_Berman	6	0.861	-1.675
kni	16	0.870	-1.68
scalloped	15	0.906	-1.697
mad	8	0.914	-1.701
Bcd_Berman	12	0.925	-1.706
ftz-f1	13	0.933	-1.71
Cad_Rajewsky	12	0.951	-1.718
Pnt_Halfon	12	0.956	-1.721

Table 4.5.4 LDL. Literature derived list

Tables 4.5.* Over-representation analyses of all the compiled gene lists. Feature name indicates the name and origin of the matrix used. All matrices with no accompanying publication title were obtained from TRANSFAC through the Motifscanner utility. The number n refers to the number of times this feature appears in the active set. The Prob(occ{b}>=n) column refers to the p-value representing the probability to find even more occurrences than n in the number of base pairs used in this analysis (all the utilised CRM sequences). The SIG value refers to the significance of the over-representation of the relevant TF. Information for interpreting SIG values can be found in section 4.8.

4.8 TFBS analysis synopsis

Tables 4.5.* present the results of the TFBS analyses performed on the LDL, CCEL, FGPL and GSL collections. The number of occurrences (n) of each putative TFBS in each library is provided in addition to a p-value that represents the probability of finding even more occurrences than n in a sequence (or sequence collection in this case) of corresponding size. Van Helden et al. (2000) state that ‘‘when analyzing only one feature, a p-value smaller than 0.05 could be selected as being over-represented. However, in case of multiple features, it is better to use the SIG value (or use a Bonferroni-type family significance level)’’. In a different publication van Helden et al. (1998) state that ‘‘when selecting only the patterns for which sig >= 0, one expects less than one pattern to occur at random within each family (or enhancer library in this case). Each increment of 1 for the significance coefficient represents a drop of a factor of 10 for the occurrence probability. In other words, one expects to find at random one pattern with sig >= 1 every ten families, one with sig >= 2 every 100 families, and one with sig >= s every 10s families’’. Similarly a value of -1 signifies an expectation of finding at random 10 patterns within each family.

In light of these considerations only the over-representation of putative *giant* (*gt*) binding sites in the FGPL collection is deemed statistically significant (it generated the only positive SIG value, 0.436). *Gt* is a gap gene with a role in segmentation but it is not involved in gonadogenesis and its expression pattern doesn’t significantly overlap with that of *Six4* in the mesoderm (Mohler et al., 1989).

It is conceivable that many of the TFBSs identified in the enhancers of the genes examined in this study are functional *in vivo*. However, in the absence of supporting data (such as footprinting) it is impossible to make inferences as to the identity of these factors and the implications of their involvement in *Six4* regulation. Ways of addressing this issue are discussed in section 4.10. The following section deals with the deconstruction of the Six4-3int expression pattern through an *in vivo* enhancer partitioning assay that attempts to investigate the functional importance of the CSBs identified by the footprinting analysis described herein.

4.9 Enhancer element partitioning

The TFBS analysis of the CSBs identified in this study found two clusters of putative TFBSs in two separate CSBs (six4005F and six4007F, see Fig. 4.5.1). No other CSBs were found to harbour significant hits to TFBS matrices. Additionally another CSB (six4002F) was found to contain matches to CSBs found in the enhancers of other genes known to be linked to *Six4* (see table 4.1). In this section I attempt to separate these CSBs through an approach reminiscent of ‘‘promoter bashing’’ in order to ascertain their functional significance.

The experimental validation of the significance of the putative TF binding sites discovered within Six4-3int necessitated the isolation of these sequences in an attempt to deconstruct the mesodermal expression pattern of Six4. A partitioning and subsequent incorporation into a GFP-reporter of fragments of Six4-3int was performed in order to pinpoint the areas of Six4-3int that are responsible for patterning. To that end I decided to divide the Six4-3int enhancer in two parts, each incorporating one of the two putative TFBS containing CSBs (six4005F and six4007F) and use those to drive GFP expression in transgenic embryos. The theory behind this was that if the mesodermal expression pattern of Six4 is the product of composite modular regulation then each of the two resulting reporter constructs should recreate distinct aspects of the Six4 expression pattern allowing for the identification of the CRMs responsible for regulation. What follows is a description of the methodology used to achieve this. Fig. 4.7 illustrates the positioning of these CSBs as well as the restriction site used to partition the enhancer.

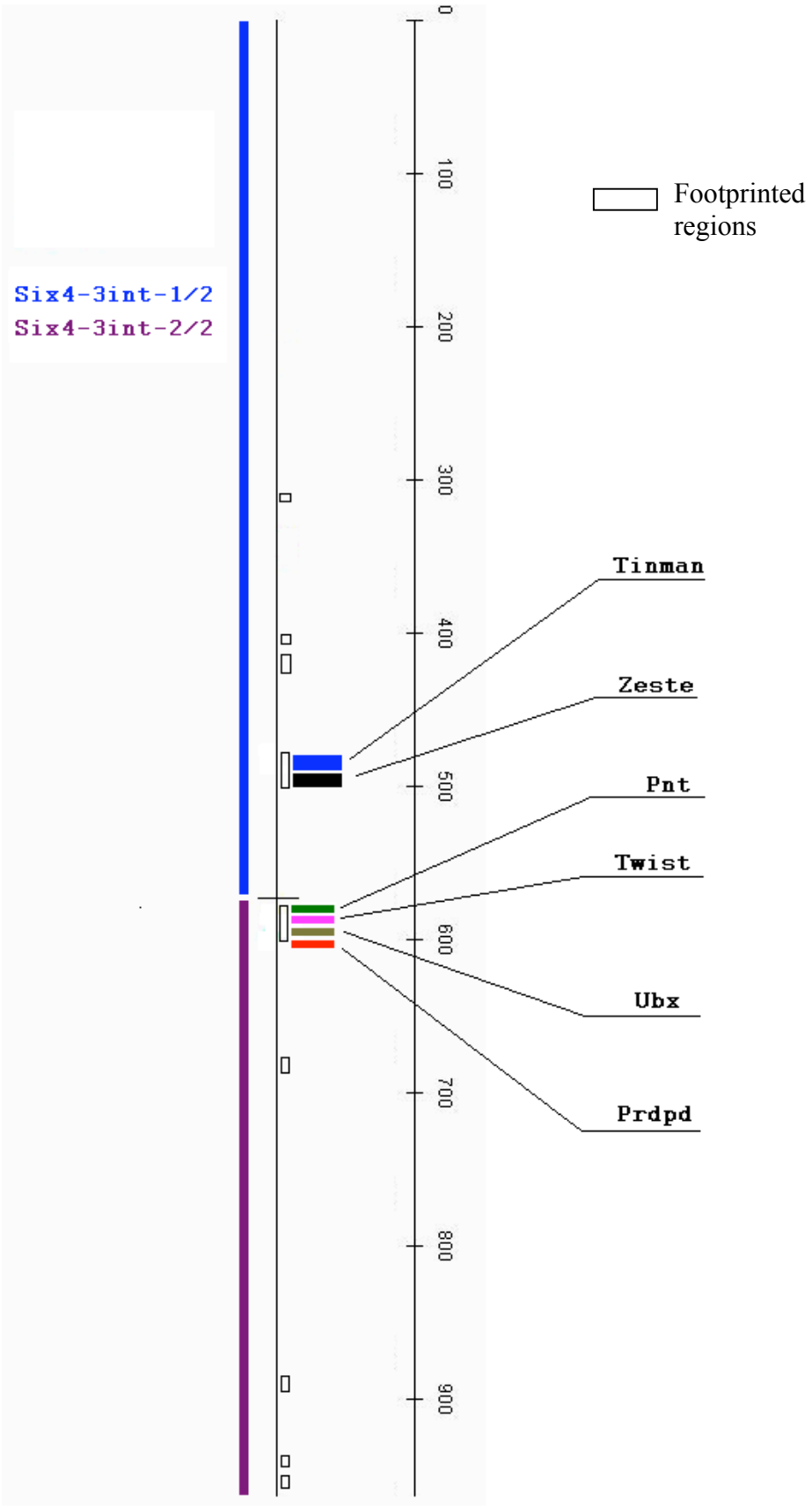


Fig. 4.7 Map of the partitioned Six4-3int. The relative positions of the CSBs thought to be implicated in *Six4* regulation are indicated. The partitioned enhancer is divided in two fragments Six4-3int-1/2 (positions 1-570) and Six4-3int-2/2 (positions 571-987) each containing one of the CSBs that were found to harbour putative TFBSs.

The primers TCTAGACAGCAAAGACCGTGATG and GGATCCGAATGGATTGCCATCCAGTTG were used to amplify the *Six4* third intron sequence from wild-type genomic DNA. The resulting PCR product was sequenced and found to deviate from the reported Six4-3int genomic sequence by 3 single base substitutions. These point mutations were found to be in non-footprinted parts of the enhancer and as such were considered to be unimportant. A restriction digest of the resulting fragment was performed using the Sph1 restriction endonuclease (restriction site at position 571 of 987) and the resulting fragments as well as the complete amplified sequence were subsequently inserted into the multiple cloning site (MCS) of the pH Stinger vector (Barolo et al., 2000). Figures 4.8 and 4.9 show the MCS of pH stinger and a restriction digest of the recombinant plasmid respectively.

The resulting plasmids were used to transform the w^{1118} πΔ2-3 expressing strain of *D.melanogaster* through microinjection. The presence of the πΔ2-3 transposase gene is expected to cause insertion of the pH-Stinger element at a random position in the *Drosophila* genome. Transformants were screened on the basis of having mosaic red eyes as a result of a transposase mediated pH-Stinger *white* gene insertion in the soma. Transformants were collected and then crossed to members of the w1118 strain to obtain a stable germ line transformant strain through the crossing out of the pΔ2-3 transposase gene. Stable transformants were screened on the basis of having red eyes as opposed to mosaic eyes caused by random excision of the white gene due to the presence of active pΔ2-3 transposase in the soma. A total of 18 transformant lines were collected but only insertions of the GFP reported gene mapped to the 2nd chromosome through chromosome marker mapping crosses were used in the following experiment.

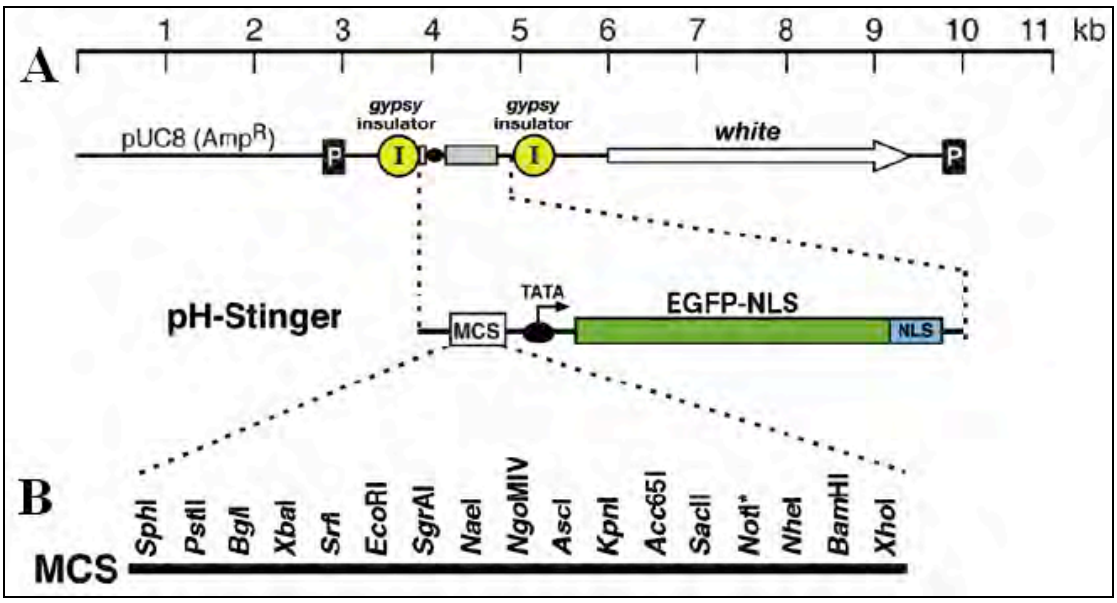


Fig. 4.8 Insulated enhanced green fluorescent protein (EGFP) vector pH-Stinger a) Diagram of GFP pH-Stinger vector. Vector uses the mini-*white* gene as a transformation marker, and contains a minimal Heat shock protein 70 (Hsp70) promoter (TATA, black oval and arrow) driving the reporter gene. Black boxes labelled P represent terminal P-element sequences required for transposition. Yellow circles labelled I represent transcriptional insulator sequences from the *gypsy* transposable element. The white rectangle represents the multiple cloning sequence (MCS). Codons for a C-terminal nuclear localization signal (NLS) from the *transformer* gene are indicated. b) Unique restriction sites in the MCS of pH-Stinger. Vector sequence and sample images are available at www.biology.ucsd.edu/labs/posakony.

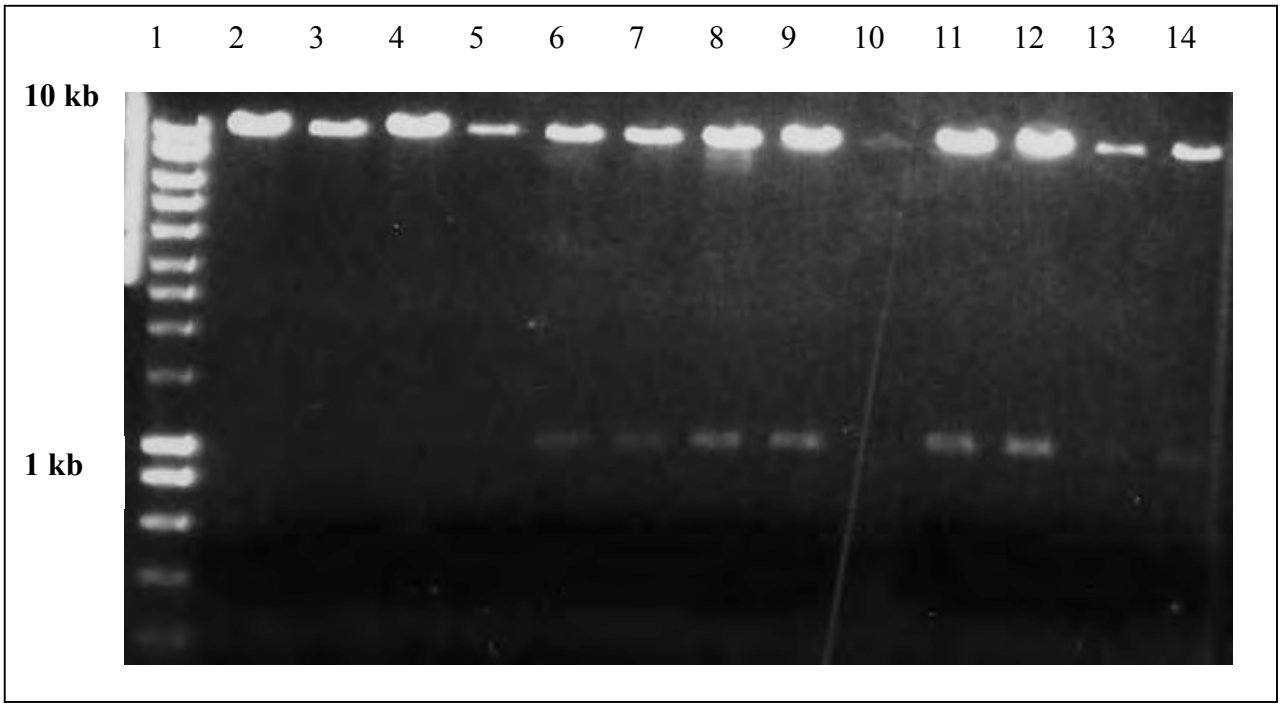


Fig. 4.9 Agarose gel electrophoresis of a restriction digest of plasmid DNA isolated from colonies grown from *E.coli* transformed with a GFP pH-Stinger + Six4-3int amplified fragment ligation. In lanes 3-14 plasmids were digested with XbaI and BamHI restriction enzymes. Restriction is expected to cause excision of the Six4-3int fragment insert. Lanes 6-9 and 11-12 show restriction fragments of a size consistent with that of Six4-3int. Lane 2 contains GFP pH-Stinger vector (no insert) digested with XbaI and BamHI.

Strains were designated Six4-3int-1/2GFP 1-4, Six4-3int-2/2GFP 1-6 and Six4-3int-GFP 1-8 based on the fragment of Six4-3int used to drive GFP expression (positions 1-570 for 1/2, 571-987 for 2/2 and 1-987 for Six4-3intGFP) and the number assigned to the different insertions. Embryos from 3 separate insertion products from all 3 groups of strains were collected at various developmental stages and were immunohistochemically labelled with a primary rabbit anti-GFP antibody (BD biosciences) and a primary anti-Eya antibody (mouse 1/100, Developmental Studies Hybridoma Bank, developed by N. Bonini) and then subsequently detected using secondary antibodies conjugated to Alexa 488 or 568 fluorochromes (Molecular Probes) for anti-GFP and anti-Eya antibodies respectively. In Figures 4.10 – 4.22 GFP is labelled green whereas Eya is labelled red. Embryo staining and microscopy was performed as described in section 5.2.3.4.

Fluorescence in Six4-3int-GFP embryos was found to be consistent with that reported by Clark et al. (2006) in all developmental stages. These findings are summarised in Figures 4.10 - 4.13. These observations serve as a positive control and corroborate the findings of these authors. For a complete description of this expression pattern see section 4.2. To confirm the Six4-3int-GFP expression pattern, embryos were also stained with a primary anti-Eya antibody and subsequently labelled with an anti-mouse antibody. The expression of Eya in stages 7-16 closely resembles that of Six4. Expression of Eya and GFP in these stages showed almost complete overlap (Figures 4.14 and 4.15)

Conversely, fluorescence in Six4-3int-GFP-1/2 embryos was found to be significantly different from that reported for Six4-3int-GFP in all insertion lines. In stages 7-9 GFP expression is detectable in mesodermal cells along the entire germ-band. However where Six4-3int-GFP expression ceases after stage 12 in all parts of the embryo apart from the SGPs, Six4-3int-GFP-1/2 expression is detected in a distinct position within each parasegment, but not in the gonad (Figures 4.16-4.19). In a dorsal view the mesodermal staining appears to consist of a bilateral pair of longitudinal bands terminating in a distinct point of fluorescence. These features are consistent with the known transient segmental characteristics of the mesoderm at this embryonic stage. It is possible that fluorescence may correspond to the myoblasts of the dorsal and lateral musculature. As mentioned previously Six4 is expressed in the somatic muscles but by stage 9 its expression is limited to the SGPs. It is unlikely that fluorescence is the result of GFP perdurance (fluorescence appears to be stronger at

stage 15, Fig. 4.16). It is more likely that Six4 expression in this part of the embryo is shut down post stage 9 through the action of TFs that bind to Six4-3int between positions 571-987. The absence of these positions would explain the inability of these factors to stop Six4 expression after stage 9. No GFP expression was detected in SGPs (see Figures 4.20 and 4.21). This fact was corroborated through the absence of Eya-GFP coexpression in the gonad (Eya is known to be expressed in the gonad at this stage). However GFP expression was detected in stages 7-16 in 2 distinct positions within each parasegment (Figures 4.15 to 4.19). These observations led me to conclude that the TFBSs responsible for SGP expression and possibly part of the mesodermal expression of Six4 are either not located between positions 1-571 of Six4-3int or are required in exerting combinatorial control in conjunction with elements present between positions 572-987. Additionally the emergence of a weak yet distinct new expression pattern hints towards the function of TFBSs in Six4-3int-1/2 whose function is either overshadowed by that of elements in Six4-3int-2/2 or held in check through the action of repressors present there. A candidate for this role is Tinman. The footprinted putative Tinman TFBS in Six4-3int-1/2 could control this expression (located within the first footprinted putative TFBS cluster identified in this study). The ectopic Six4-3int-1/2 expression is consistent with that of Tinman (dorsal mesoderm), potentially hinting towards Tinman regulation of *Six4* in at least some parts of the mesoderm being held in check through the action of other factors that bind within the Six4-3int-2/2 enhancer. The loss of this enhancer may give rise to the ectopic expression reported here. The potential involvement of Tinman in *Six4* expression is revisited in section 4.10.1.

Conversely, expression of GFP driven by Six43int-2/2 was found to be similar to that of Six4-3int-GFP after stage 13 but did not mimic the complete Six4-3int-GFP expression pattern. Expression in the SGPs in developmental stages 14-16 was found to echo that of Six4-3int-GFP (Figures 4.23 and 4.24) almost completely. Expression prior to stage 13 was either absent of severely diminished (Fig. 4.22). This phenomenon was observed in all insertion lines and as such is unlikely to be a result of insertion in a transcriptionally silent region of the *Drosophila* genome.

Based on the observations made during the enhancer element partitioning analysis it was postulated that the majority of the TF binding sequences (TFbss) responsible for conferring the Six4-3int expression pattern in the early mesoderm are located within the first part of Six4-3int (positions 1-570) and the entirety of the TFBSs

responsible for Six4 expression in the SGPs are located within the second half of Six4-3int (positions 571-987). Refinement of the Six4-3int expression pattern is mediated by the action of repressor TFBSs present in both fragments of the enhancer. The implications of this with respect to the validity of the putative TFBSs identified previously will be discussed in section 4.10.

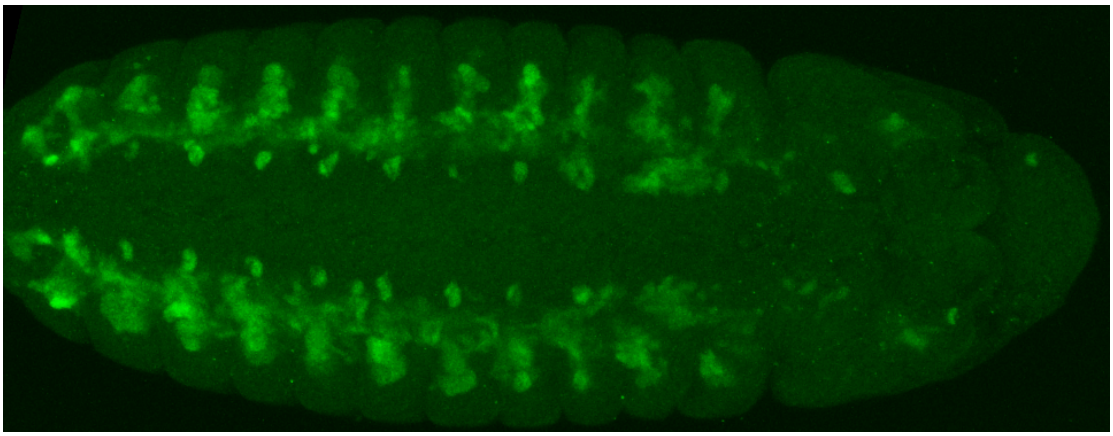


Fig. 4.10 Expression of Six4-3int-GFP (insertion line 3) developmental stage 9 (dorsal view). GFP expression is segmental and spans the entirety of the mesoderm.

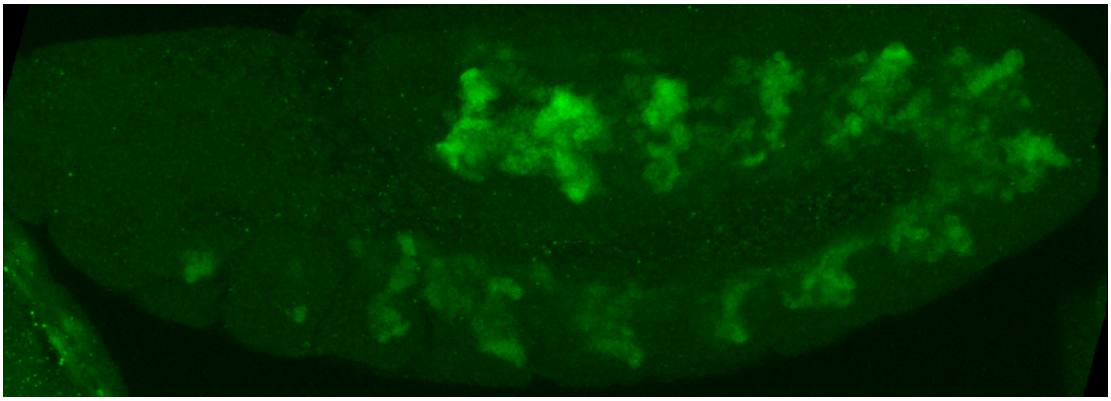


Fig. 4.11 Expression of Six4-3int-GFP (Insertion line 7), developmental stage 9 (lateral view).

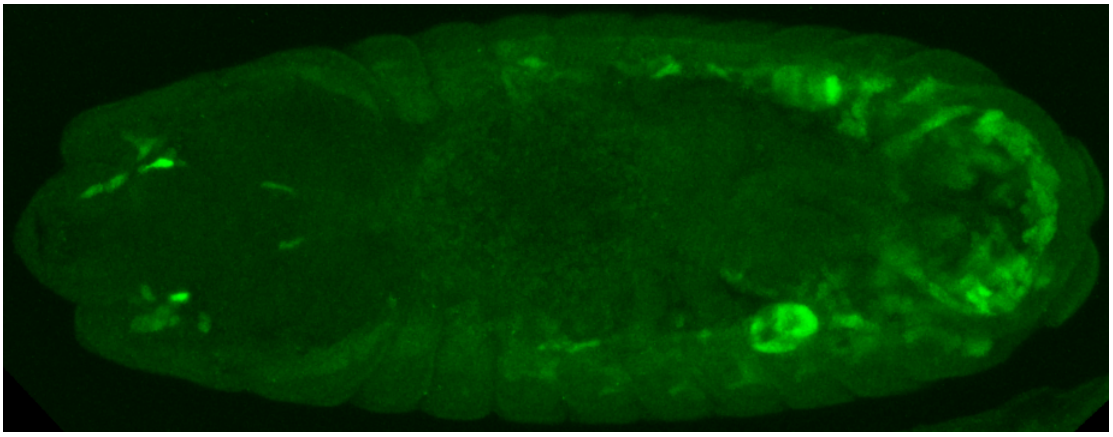


Fig. 4.12 Expression of Six4-3int-GFP (Insertion line 3), developmental stage 16 (gonadal expression, dorsal view). By stage 15 GFP expression is refined and can only be seen in the gonads (dorsal view).

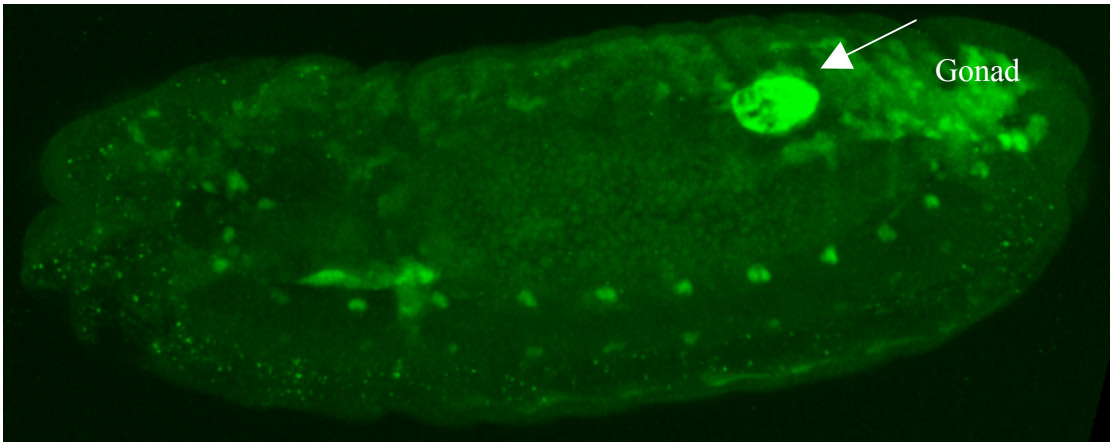


Fig. 4.13 Expression of Six4-3int-GFP (Insertion line 7) developmental stage 16 (lateral view)

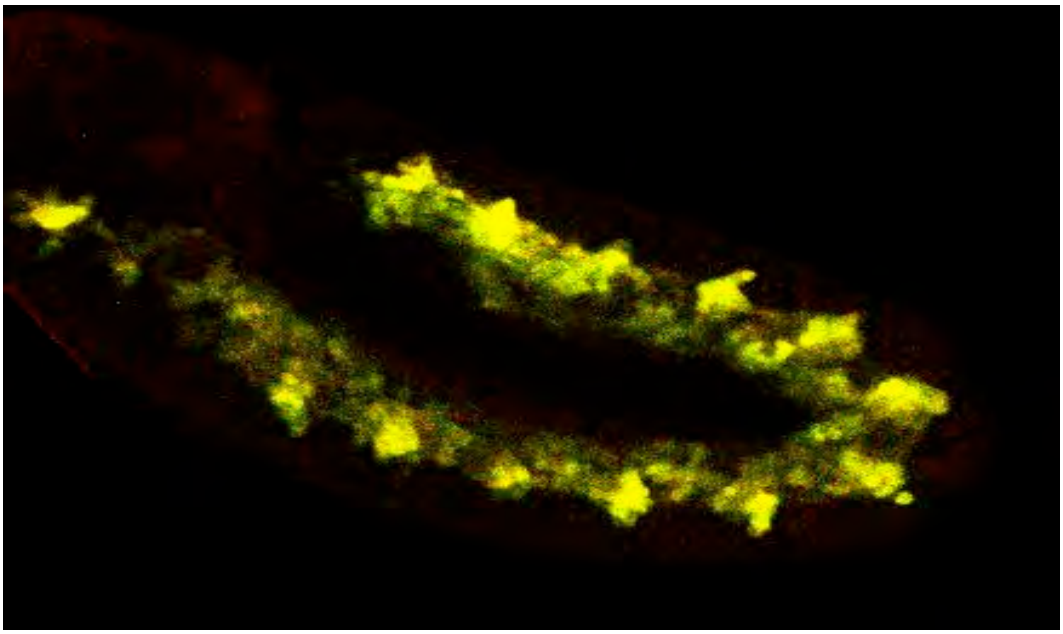


Fig. 4.14 Expression of Six4-3int-GFP (insertion line 1) developmental stage 9. GFP expression shows complete overlap with that of Eya (yellow fluorescence) (lateral view).

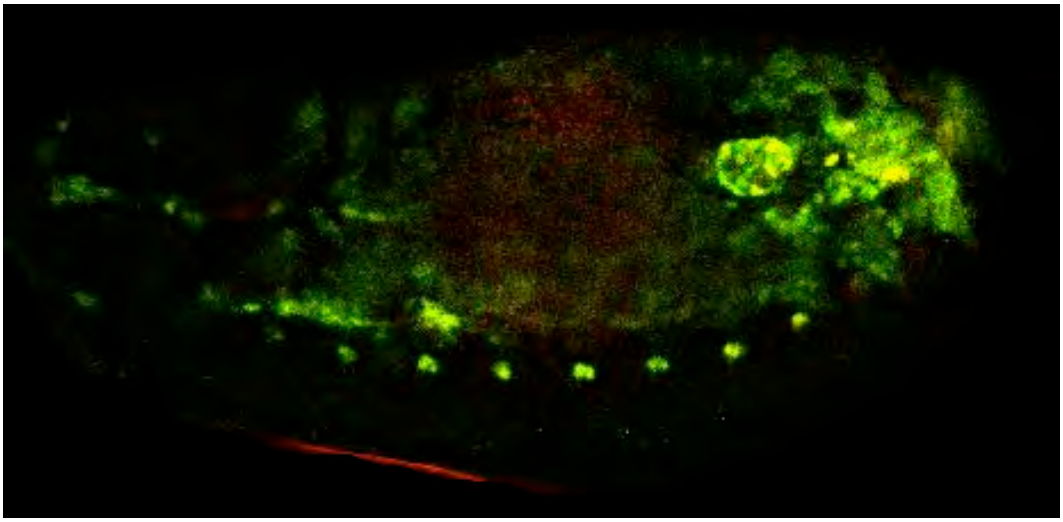


Fig. 4.15 Expression of Six4-3int-GFP (insertion line 1) developmental stage 16. GFP expression shows complete overlap with that of Eya (yellow fluorescence) (lateral view).

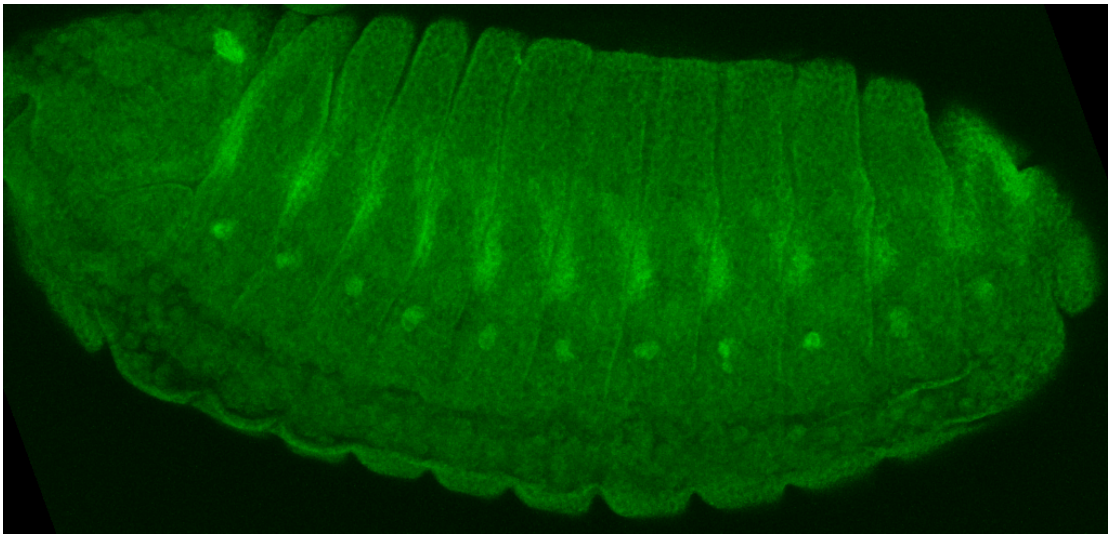


Fig. 4.16 Expression of Six4-3int-1/2GFP (insertion line 1) developmental stage 15. GFP expression is segmental and occurs in pattern resembling that of the myoblasts of the dorsal and lateral musculature.

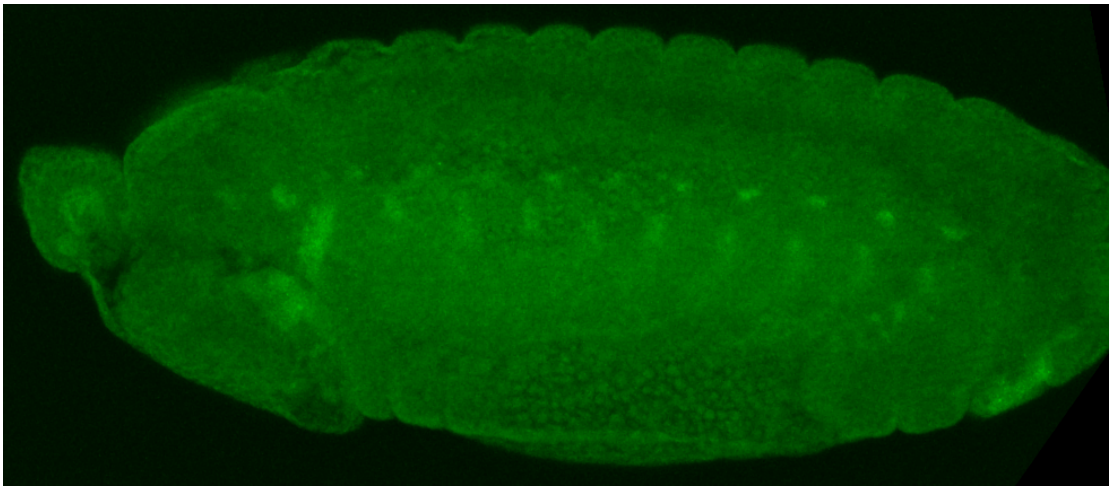


Fig. 4.17 Expression of Six4-3int-1/2GFP (insertion line 4) developmental stage 12. Expression in what could be the dorsal musculature is starting to become apparent.

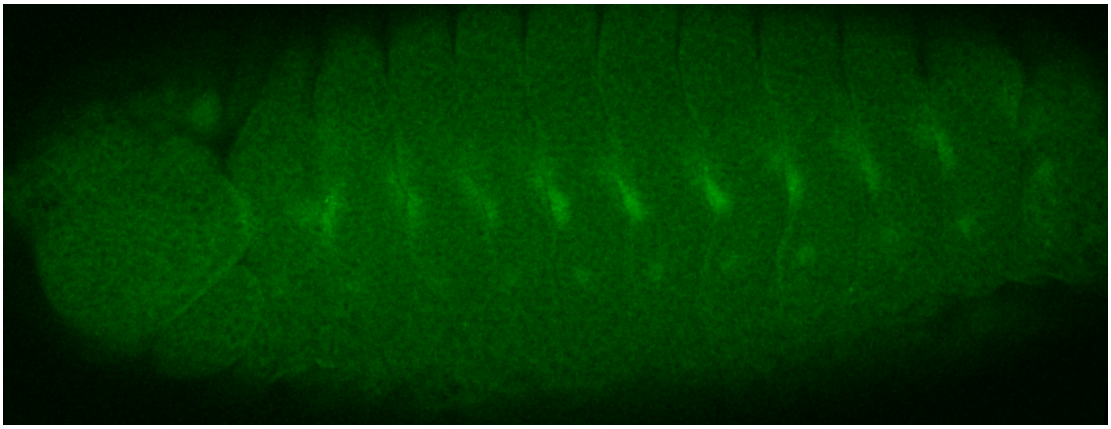


Fig. 4.18 Expression of Six4-3int-1/2GFP (Insertion line 4) developmental stage 15

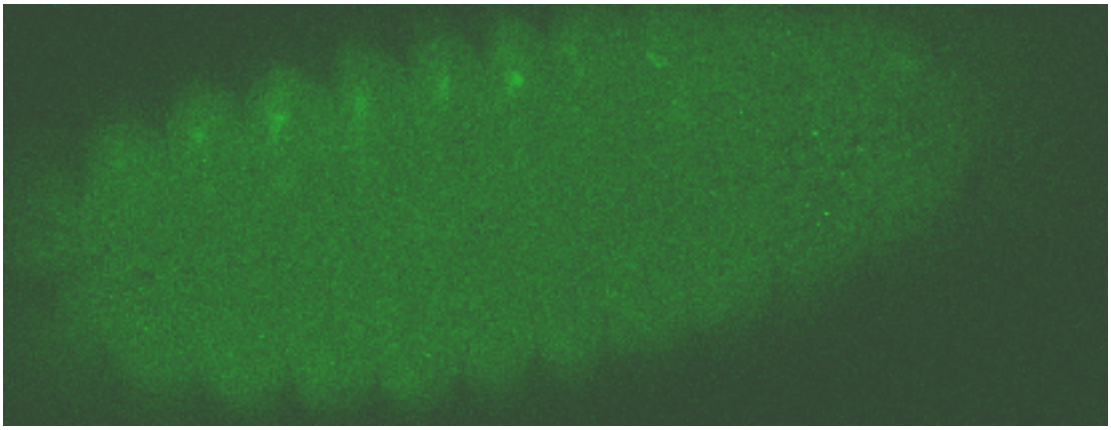


Fig. 4.19 Expression of Six4-3int-1/2GFP (Insertion line 2) developmental stage 16. No expression in the gonad is visible.

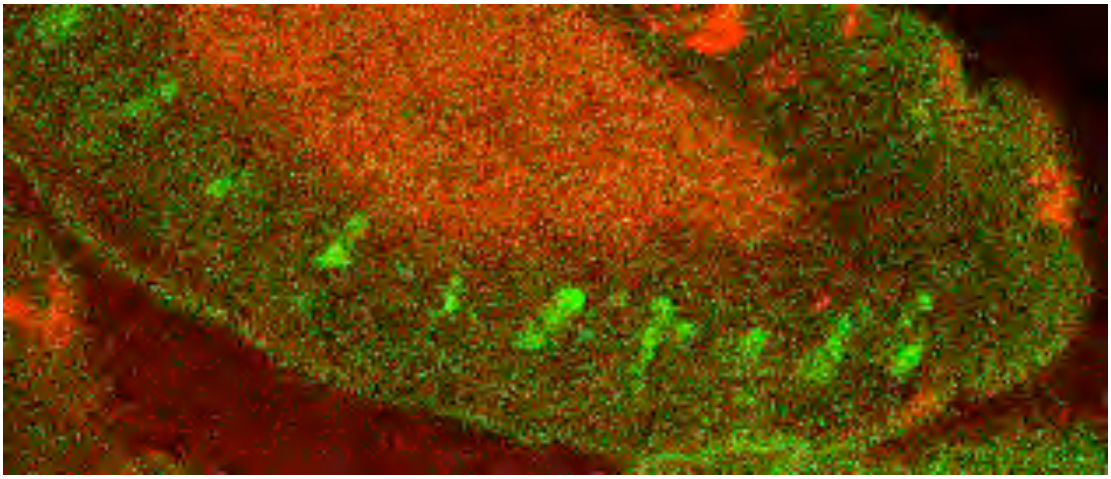


Fig. 4.20 Expression of Six4-3int-1/2GFP (insertion line 4) developmental stage 16. Expression of GFPshows little overlap with that of Eya (strongly expressed in the gonad). No expression in the gonad is visible.

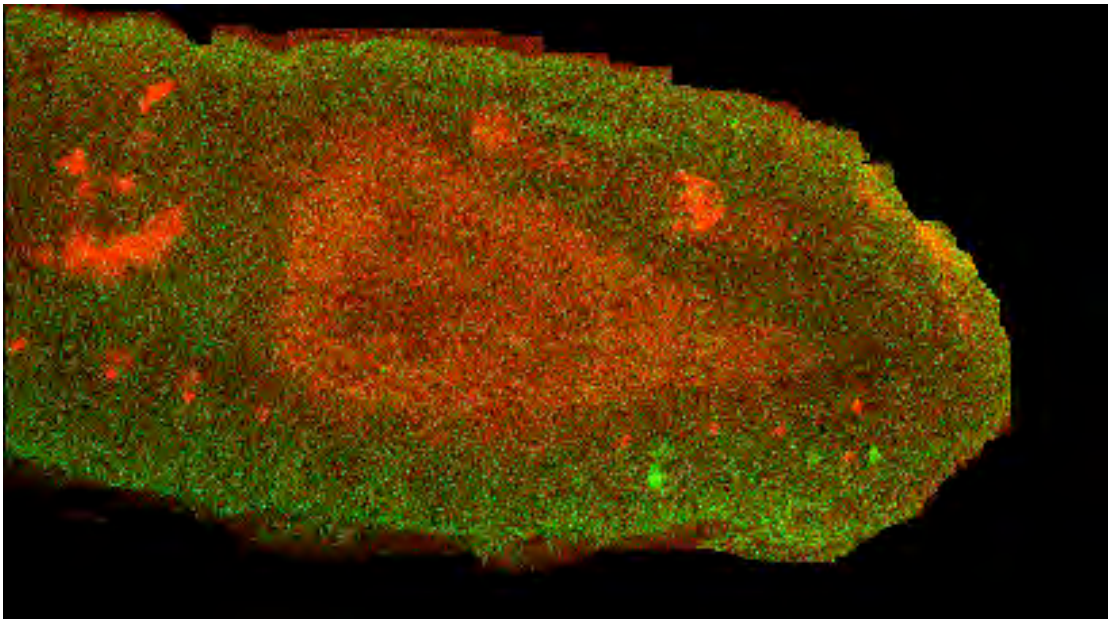


Fig. 4.21 Expression of Six4-3int-1/2GFP (insertion line3) developmental stage 16. Expression of GFP shows little overlap with that of Eya (strongly expressed in the gonad). No expression in the gonad is visible. By contrast Eya expression in the gonad is clearly visible.

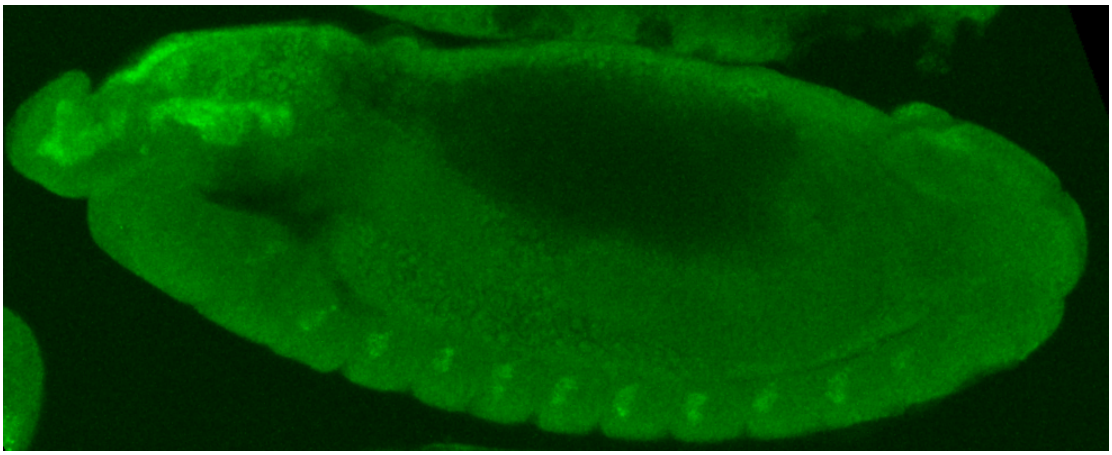


Fig. 4.22 Expression Six4-3int-2/2GFP (Insertion line 2) developmental stage 13. Little mesodermal expression is visible.

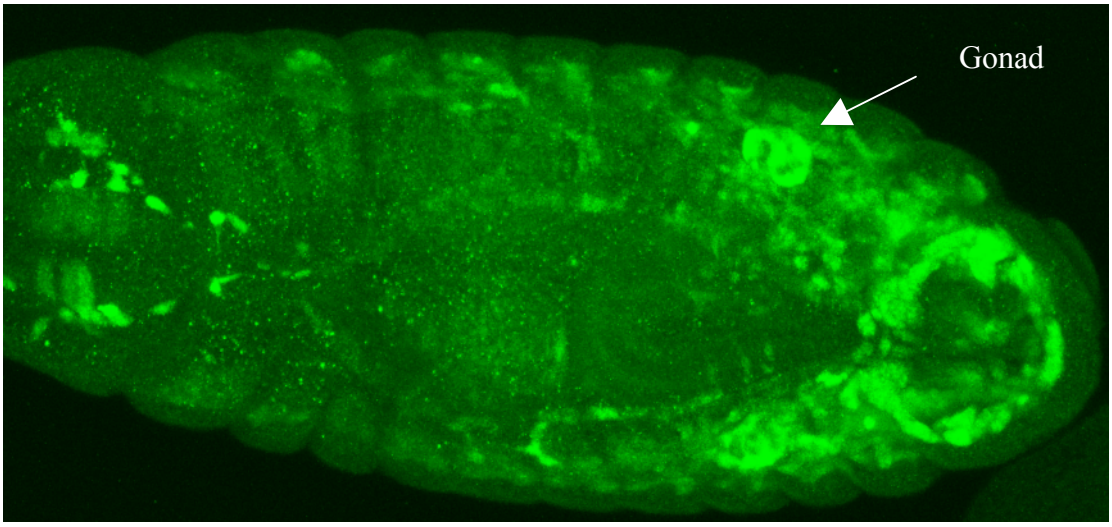


Fig. 4.23 Expression of Six4-3int-2/2GFP (Insertion line 2) developmental stage 16. Strong expression can be seen in the coalesced gonads.

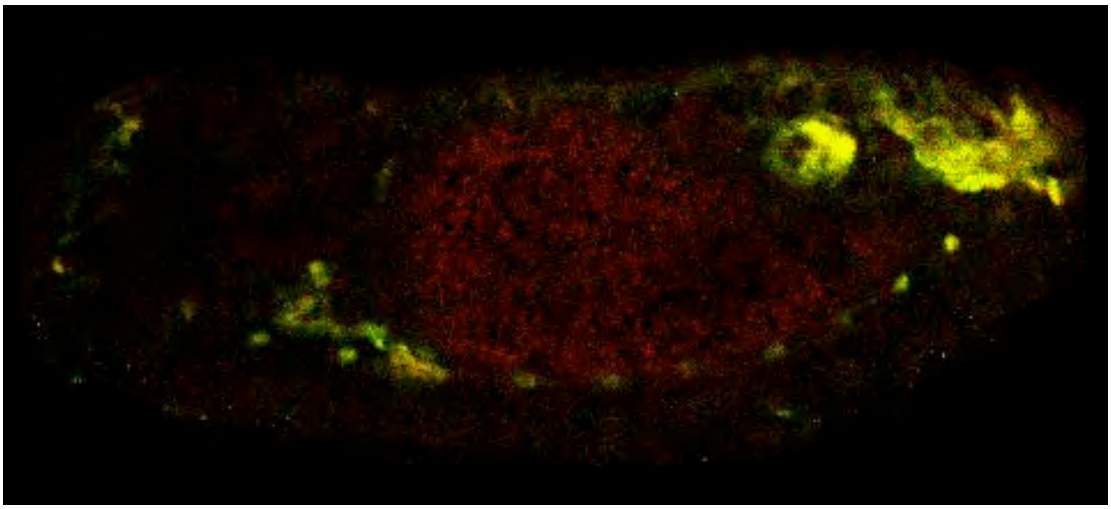


Fig. 4.24 Expression of Six4-3int-2/2GFP developmental stage 16. Expression of GFP shows complete expression with that of Eya (yellow).

4.10 Discussion

This study has attempted to elucidate the regulation of *Six4* and define its place in the tapestry of interactions that underpin mesodermal development. Using an unbiased *in silico* approach I have identified a number of putative TFBSs within the *Six4* 3rd intron enhancer element known to be responsible for the mesodermal expression of *Six4*. Some of these sites may contribute to conferring a Six4-like mesodermal expression pattern. This possibility has been investigated by testing for overrepresentation of these factors in the putative (or occasionally known) enhancers of other genes known to either be expressed in a *Six4* like pattern, or to be linked to *Six4* through interacting developmental pathways. The modular nature of CRMs in

conferring distinct expression patterns is well known and has led me to believe that a distinct group of TFBSs are required to allow for expression in a pattern consistent with that of *Six4*. In order to discover this pattern I have identified a number of genes expressed in a pattern reminiscent of that of *Six4*. The rationale behind this is that co-expressed genes are likely to be under the control of similar CRMs. Expression data available through a number of sources was used in identifying such genes and extracting their putative CRMs.

The TFBSs analyses described in this chapter have failed to identify over-represented TFBSs in the putative enhancers of co-expressed genes. The statistical significance of any findings stemming from TFBS analyses of these sequences is inadequate to support any conclusions on the identity of the factors that confer the *Six4* mesodermal expression pattern. A number of factors may be responsible for this.

Firstly, the gene lists scanned for these elements are not all inclusive (as can be expected from any list constructed of *Drosophila* genes) since expression data is unavailable for most of the entries in Flybase. This forces the researcher to work with more limited gene lists that generate information of lesser statistical significance. More importantly, and depending on the nature of the list, many genes may be included in the collections through co-expression by potentially distinct pathways (this is truer of the FGPL list) and will therefore be controlled by different factors. I do not expect this factor to greatly influence this analysis since the expression pattern in question (expression in the gonadal mesoderm) is fairly specialised and thus unlikely to be the result of divergent regulatory pathways. I took steps towards minimising this effect even further by compiling a list of genes with nigh identical expression patterns (the CCEL). However, this compilation depends greatly on the in situ images available through the BDGP, the quality of which has been known to vary greatly.

More importantly, if one is to extract the regulatory information out of these gene lists then putative CRMs need to be identified. Whilst some of these are already known and a reasonably reliable guess can be made about the location of the others it is uncertain that all the real CRMs of the genes in question have been used in this analysis. I have argued in chapter 3 that most regulatory information is often contained within 1kb of the transcriptional start. Whilst this holds true, this approach is much more suited to a genome-wide matrix scan that is permissive towards the inevitable loss of some regulatory information. A TFBSs co-regulation analysis is heavily dependant on the identification of real CRMs and will therefore suffer if

CRMs are missed. This in turn can easily happen if genes are controlled by CRMs that lie outside the searched space (as is the case of *Six4* where the mesodermal enhancer is located in an intron). Methods do exist for determining the location of CRMs (see Rajewsky et al., 2002; Johansson et al., 2003; Alkema et al., 2004; Sharan et al., 2003; Grad et al., 2004; Aerts et al., 2003; Aerts et al., 2004; Sharan et al., 2004; Hu et al., 2008; Xie et al., 2008 amongst others). However a detailed analysis of the identified genes for putative CRMs lies beyond the scope of this study (as attested by the extensive amount of reported methodology on this subject). The compiled gene lists presented herein would provide a very good starting point for such an analysis.

I have chosen to focus this study on the putative TFBS reported within phylogenetically footprinted regions of the Six4-3int enhancer. Eleven such conserved sequence blocks were identified in the Six4-3int enhancer. These footprints harboured putative TFBSs for six different TFs arranged in 2 clusters. Four of these TFs are expressed in the mesoderm and are suspected of regulating Six4. Additionally, these footprints contained matches to footprints in the enhancers of nine mesodermally expressed genes. Two of these genes have previously been associated with Six4 (*tinman* and *serpent*). Interestingly, these findings indicate that Tinman is potentially a regulator of *Six4* whilst being subject to the same regulatory signals that confer Six4 mesodermal expression. Since the conclusion of this study Eileen Furlong's group has shown though ChIP that Tinman binds to the Six4-3int enhancer, possibly through the reported foorprinted BS (Eileen Furlong, personal communication). This finding lends credence to other suspected associations highlighted in this study (section 4.10.1). Another likely regulator of *Six4* is the mesodermal activator Twist, a putative TFBS of which is also footprinted in Six4-3int.

Finally, I have tested the impact of these putative TFBSs on regulation by performing an expression pattern partitioning analysis on the Six4-3int enhancer. By dividing the enhancer into two parts and using those to drive reporter constructs I have tried to decipher the modular regulatory control exerted on Six4. The enhancer partitioning effectively segregates the footprinted putative TFBSs of Tinman (Six4-3int1/2) and Twist (Six4-3int2/2).

The results of the enhancer partitioning analysis suggest that the regulatory signals that confer mesodermal expression up till stage 9 originate from the first part of the enhancer (the one containing the putative BS) since this enhancer is sufficient for driving expression in the early mesoderm but not the SGPs (no expression in the gonad). Conversely, expression in the SGPs is predominantly conferred by the second

part of the enhancer that is sufficient for driving gonad expression but cannot account for early mesodermal expression. These results can provide the basis for more restrictive enhancer partitioning or mutagenesis analyses (targeting specific CSBs).

Given the number of TFs that are likely to influence *Six4* regulation I will refrain from speculating on the implications of this analysis on the potential involvement of specific TFs. The only factor that is currently known to be a very likely to be a *Six4* regulator is Tinman and the nature of its putative association with *Six4* will be discussed herein.

4.10.1 The potential interaction between Tinman, Six4 and Twist

In some aspects of somatic muscle patterning, the roles of Tinman and Six4 appear complementary (Clark et al., 2006). Embryos lacking Tinman function also show loss of SGPs and mild defects in fat body development (Boyle et al., 1997; Moore et al., 1998_1), suggesting that the transient pan-mesodermal expression of *tin* has a role to play in these cells as well as in some ventral and lateral somatic muscles. Clark et al. (2006) investigated the possibility that regulation of *Six4* may account for these *tin* phenotypes. In *tin* mutant embryos, a marked reduction in Six4-III-GFP (Six4-3int-GFP) expression in most parasegments was observed by these authors. It was therefore expected that *Six4* was under direct or indirect Tin regulation. This loss of expression was thought to partly underlie the SGP and fat body defects in *tin* mutants, although the remaining Six4 expression was thought to be sufficient for the Tin-independent somatic muscles that required Six4 and which were unaffected in those mutants.

However the present study shows that GFP driven by the part of the enhancer lacking the putative Tin TFBS is expressed in the SGPs. Factors that could account for this are the possible presence of another Tinman TFBS in Six4-3int-2/2 (none reported) or further indirect regulation of *Six4* by Tinman through other factors. Of these two explanations the second seems the most likely. Irrespectively of this the removal of positions 571 onwards from the Six4-3int enhancer gives rise to an expression pattern reminiscent of Tinman expression. This observation leads me to believe that Tinman-mediated regulation of *Six4* is potentially modulated by elements that lie after position 571. Furthermore the findings of this study suggest that if no other Tinman TFBS exists within the Six4-3int-2/2 enhancer then expression of *Six4* in the SGPs is the product of indirect Tinman regulation.

Finally, *tinman* itself is known to be regulated by Twist (this regulation is modulated by the binding of Eve although Twist is sufficient for ectopic Tinman expression, Yin et al., 1997) a potential regulator of *Six4* itself through a footprinted putative TFBS within the Six4-3-int-1/2 enhancer. The absence of the early mesodermal GFP expression in Six4-3int-2/2 could be a result of the loss of the putative Twist TFBS located in the footprinted putative TFBS cluster identified in Six4-3-int-1/2. Twist is a transcription factor that is responsible for most of the expression in the mesoderm. It is likely that *Six4* expression in the mesoderm is switched on by Twist and then later refined to the SGPs through the action of Tinman, probably in association with some other, as yet unknown, factor.

The potential regulation of *Six4* by Tinman also strengthens the possibility of involvement of another factor in *Six4* regulation. Tinman action is usually mediated by co-regulators in specifying various structures including the heart, visceral musculature, and dorsal body wall muscles (Azpiazu et al., 1996) and could act in a similar fashion in its regulation of *Six4*. The following section addresses ways of testing this theory.

4.10.2 Future experiments

The experiments described herein, whilst not exhaustive, provide a basis for elucidating *Six4* regulation even further. As discussed earlier, the CRM analysis of *Six4* co-regulated genes is not exhaustive and can be explored further. A combinatorial CRM analysis (see section 4.10 for references to methodology) can yield better results for detecting putative CRMs of co-expressed genes. These CRMs can then be subjected to a footprint analysis thus isolating CSBs from all co-expressed genes. Such an approach would not only drastically reduce the search space for TFBSs prediction algorithms thus removing potential masking effects that can hamper over-expression analyses but will also generate more dependable data upon which further analyses can be based.

Additionally, the putative TFBSs identified within CSBs in Six4-3int (as well as the conserved mesodermal sequences) can be subjected to partitioning or mutagenesis (either independently or jointly) thus providing definitive evidence concerning the utility of these elements. This approach can be equally applied to potential Six4 target sequences (Chapter 3) and to putative TFBSs identified in the Six4-3int enhancer as well as the putative regulatory elements identified herein. The mutagenesis of the

putative Tinman and Twist TFBSs in particular would be of great interest as it would test the hypotheses outlined in section 4.10.1.

Chapter 5 – Materials and Methods

Materials and Methods

5.1 Materials

5.1.1 Media

5.1.1.1 Bacterial media

Luria Bertani Broth (LB)

Bacto tryptone (Difco), 10g; Bacto Yeast extract (Difco), 5g; NaCl, 5g; per litre adjusted to pH7.2

Luria Agar (L-agar)

Luria broth with 15g/l Bacto agar (Difco). Ampicillin (Penbritin, Beecham Research) was added to LB and L-agar to a final concentration of 100μg/ml where indicated

SOC Buffer

LB with 3.6 g/l glucose, 0.1 $MgSO_4$ and 0.1 $MgCl_2$

2x TY Broth

Bacto tryptone (Difco), 16g; Bacto Yeast extract(Difco), 10g; NaCl, 10g; per litre adjusted to pH 7.4

5.1.1.2 *Drosophila* media

'French' fly food

Oxoid No.3 agar, 7.5g; polenta, 55g; dried flake yeast, 550g; nipagen (150mg/ml made up in 95% ethanol), 10ml dH_2O, 100ml

Dundee Fly Food

443g brewers yeast, 714g maize, 57g live yeast, 786g glucose, 27g nipegin, 107g agar 32ml propionic acid up to 10L with water.

Grape juice agar

Bacto agar (Difco), 205g per 100ml pure grape juice

5.1.2 Materials

5.1.2.1 Chemicals

Chemicals where purchased from Fisher, New England Biolabs, Promega, Roche, Sigma, Stratagene, BDH and Boehringer Mannheim

5.1.2.2 Solutions

TE

10mM Tris; 50mM EDTA; adjusted to pH8

PEG solution

50mM Na_2HPO_4; 22mM KH_2PO_4; 86 mM NaCl; 1mM $MgSO_4$; 0.1 mM $CaCl_2$; 0.001% gelatine

4 x Agarose gel loading buffer

20% glycerol(v/v); 0.05% bromophenol blue in TE

TAE

45 mM Tris-borate; 1mM EDTA

PBS

137 mM NaCl; 2.68 mM KCl; 10 mM Na_2HPO_4; 1.76mM KH_2PO_4 pH7.4

IPTG 100 mM

23.8 mg IPTG was dissolved in 1 mL water. Solution was filter sterilized and stored at 20^o C.

5.1.2.3 Enzymes

Restriction enzymes were purchased from New England biolabs and/or Promega

5.1.2.4 Radioactive Isotopes

A-^{32}P-dCTP (3000Ci/mM) was supplied by Amersham

5.1.2.5 Plasmids

Plasmid name	Use	Source
pGEM-T(easy)	TA cloning vector for cloning PCR products	Promega
hStinger	transgenic_transposon	FlyBase, (Barolo et al., 2000). FlyBase inference based on genome sequence analysis.
TOPO	cloning vector for cloning PCR products	Invitrogen
P-GEX-2T	GST recombinant protein generation	GE healthcare

5.1.2.6 Oligonucleotides

Oligonucleotide denomination	Sequence	Use
R76 random core oligonucleotide	5’-gtcagatctcttggcattn26actgtcgatgcggc actgtc -3′	Initial SELEX oligo
R63 random core oligonucleotide	5’-gttagacagtgccgcatcgacagtN_{13}cga acgcaatgccaagagatctgac-3’	Revised SELEX oligo
R60 random core oligonucleotide	5’-gttagacagtgccgcatcgacagtN_{10}cga acgcaatgccaagagatctgac-3’	Revised SELEX oligo
R57 random core oligonucleotide	5’-gttagacagtgccgcatcgacagtN_{7}cga acgcaatgccaagagatctgac-3’	Revised SELEX oligo
M13/pUC seq. primer (-20), 17-mer	5'-d(gtaaaacgacggccagt)-3'	Sequencing primer

RAmpPrimer1	5’-catatgttctccacggatcagatccagtgc-3’	Revised SIX and homeodomain amplification primer 1
RAmpPrimer2	5’-aagcttcagcaccgacatgatgtccgg-3’	Revised SIX and homeodomain amplification primer 2
AmpPrimer1	5’-tctagacagcaaagaccgtgagttg -3’	SIX and homeodomain amplification primer 1
AmpPrimer2	5’-ggatccgaatggattgccatccagttg-3’	SIX and homeodomain amplification primer 2
core1	5’-cgttagacagtgccgcatcgacagtgacctg acgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
core2	5’-cgttagacagtgccgcatcgacagt agcctgacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
core3	5’-cgttagacagtgccgcatcgacagt aatctgacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
core4	5’-cgttagacagtgccgcatcgacagt aacgtgacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
core5	5’-cgttagacagtgccgcatcgacagt aacttgacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
core6	5’-cgttagacagtgccgcatcgacagt aaccagacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
core7	5’-cgttagacagtgccgcatcgacagt aacctaacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
core8	5’-cgttagacagtgccgcatcgacagtaacct ggcgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
opti9	5’-cgttagacagtgccgcatcgacagtaacct gccgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo

opti10	5’-cgttagacagtgccgcatcgacagtcacct gacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
opti11	5’-cgttagacagtgccgcatcgacagtaacat gacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
opti12	5’-cgttagacagtgccgcatcgacagtaaccc gacgaacgcaatgccaagagatctgac-3’	Mutagenic Six4 consensus sequence oligo
opti13	5’-cgttagacagtgccgcatcgacagtaaccg gacgaacgcaatgccaagagat ctgac-3’	Mutagenic Six4 consensus sequence oligo
Primer1	5’- gacagatctcttggcattgcgttcgt -3’	Random core oligo amplification primer 1
Primer2	5’- cgttagacagtgccgcatcgacagt-3’	Random core oligo amplification primer 2
ARE1	5’-gatccC**CGGTGTCAGGTTGC**TCCGGTAACGGTGACGTGCg-3’	Six5 binding sequence
ARE2	5’-cGGCACGTCACCGTTACCGGA**GCAACCTGACACCG**Ggatc-3’	Six5 binding sequence

the central Nx represents x random oligonucleotides based on equal incorporation of A, G, C, and T at each position.

5.1.2.7 *E.coli* strains

Name	Genotype and use	Reference
DH5α	deoR, *endA1*, *gyrA96, hsdR17* ($r_k^- m_k^+$), *supE44, thi-1, rec A1, relA, Δ(lacZYA-argF)U169, deoR(φ80δlacZΔM15), F⁻, λ⁻*	Hanahan, 1983
BL21	F-, *omp*T, *hsd*S (r-B, m-B), *gal*, *dcm*	Wood, 1966

XL1-blue	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac* [F´ *proAB lacI*qZΔ*M15* Tn*10* (Tetr)]	Stratagene

Genes listed signify mutant alleles. Genes on the F´ episome, however, are wild-type unless indicated otherwise

5.1.2.8 *Drosophila* melanogaster strains

Name (genotype)	**Reference/Source**
πΔ 2-3	Lab stock
W1118, w-	Lab stock
Oregan R, wild type strain	Lab stock

5.2 Methods

5.2.1 Manipulation of bacteria

5.2.1.1 Growth of *E.coli* cultures

E.coli cultures were grown by inoculation of bacteria from a single colony into LB or 2xTY broth and incubation for 14-16 hours at 37^{o} C with aeration by vigorous shaking. For strains carrying ampicillin resistant plasmids, LB was supplemented with ampicillin.

5.2.1.2 Storage of *E.coli* cultures

For long term storage of *E.coli* cultures in logarithmic phase growth were mixed an equal volume of glycerol, placed in sterile tubes and kept at -70^{o} C. To grow bacteria from frozen culture a small portion was removed using a sterile loop and streaked on an L-agar plate, with ampicillin if required.

For short term storage of up to six weeks, bacteria were streaked onto agar plates which were incubated at 37^{o} C for 14-16 hours for colony growth and then kept at 4^{o} C.

5.2.1.3 Transformation of bacteria

Transformation of *E.coli* by ligation products, or when a high transformation efficiency was required (>$3x10^8$), was carried out by electroporation according to Heery et al. (1989). Single colonies were grown in for 16h in 15 ml of 2xTY medium and cells were harvested by centrifugation at 5000 r.p.m. for 10 minutes at 4^oC. Cells were washed by resuspension in 50ml of ice-cold Milli-Q® Ultrapure Water and subsequently collected by centrifugation. This step was repeated thrice and cells grown from a single colony were resuspended in a final volume of 140µl of Milli-Q H^2O. 40 µl of cell suspension were mixed with 1 µl of DNA solution and transferred to an electroporation cuvette (0.2cm, Bio-Rad). A single pulse at 2.5 kV, 15 µF, 200Ω was applied. 1 ml SOC was added immediately and the mixture transferred to a culture tube. Cells were then incubated at 37^o with shaking for 40 minutes. Several dilutions of SOC buffer were made and plated on L-agar with ampicillin. For selection of inactivation of β galactosidase expression, 100 µl of 100mM IPTG and 20µl 50mg/ml X-gal were spread onto the plates which were then incubated for 30 min at 37^o for absorption prior to use.

Alternatively for transformation of *E.coli* by all other plasmids ‘’Ultra-Competent’’ cells were prepared as described by Inoue et al. (1990) and transformation was carried out using the heat-shock method..

5.2.2 *In vitro* manipulation of DNA

5.2.2.1 Small scale preparation of plasmid DNA

Small scale preparation of plasmid DNA from *E.coli* cultures was carried out using the Wizard® *Plus* SV miniprep DNA purification system (promega) according to the manufacturer’s instructions. This method involves alkaline lysis of bacteria followed by a brief treatment with alkaline protease to inactivate endonucleasese released on cell lysis. Plasmid DNA is then purified by binding to a column, washing in a 60% ethanol solution to remove impurities and finally elution in dH_2O.

5.2.2.2 Large scale preparation of plasmid DNA

Preparation of up to 100µg of plasmid DNA from *E.coli* cultures was carried out using the Qiagen plasmid midi kit (Qiagen GmbH and Qiagen Inc) according to the

manufacturer's directions. This method is similar to the miniprep method described previously. Alkaline lysis of *E.coli* is followed by binding of plasmid DNA to an anion exchange resin under low salt and pH conditions. The resin is washed in a medium salt buffer, and the DNA eluted by high salt. Finally the DNA is concentrated by isopropanol precipitation.

5.2.2.3 Large scale preparation of plasmid DNA for injections

Liquid bacterial cultures were transferred to 50ml Falcon tubes and centrifuged at 1000rpm for 20 minutes at 4ºC. The pellets were drained thoroughly and resuspended carefully using a pastette in 2ml of solution 1 (50mM Glucose, 25mM Tris pH 8, 10mM EDTA, 5mg/ml lysozyme, prepared just before use) per 50ml of culture and left at room temperature for 10 minutes. 4ml of solution 2 (0.2 M NaOH, 1% SDS- prepared just before use) was added and the solution was mixed thoroughly but not vigorously. The viscous mixture was incubated on ice for 10 minutes with regular gentle agitation. 3ml of solution 3 (3M KOAc, 1.3M HCOOH) were added with immediate, thorough mixing and placed on ice for 15 minutes. The mixture was centrifuged at 4500rpm for 15 minutes. The clear supernatant was transferred to a clean tube avoiding transfer of any precipitate. 0.6 (v/v) of 100% isopropanol was added and the solution was mixed and incubated at room temperature for 5 minutes. The tube was then centrifuged at 4,500rpm for 10 minutes. The supernatant was discarded and the pellet rinsed with 2ml of 70% ethanol. The inner walls of the tubes were wiped clean and the still wet pellet dissolved in 1ml TE. The DNA solution was transferred to Eppendorf microfuge tubes and placed on ice for 5-10 minutes. An equal volume of cold 5M LiCl (stored at -20ºC) was added and the tubes were incubated on ice for 5 minutes, followed by centrifugation at 14,000rpm for 5 minutes. The supernatant was transferred to clean Eppendorf tubes (on ice) and an equal volume of isopropanol was added. The tubes were incubated on ice for 10 minutes and then centrifuged at 14,000rpm for 5 minutes. The supernatant was discarded and the pellets air-dried at RT. The pellets were then resuspended in a total of 300µl TE.

To remove RNA, 1µl DNAse-free RNAse (10mg/ml stock) was added and the mixture incubated at 37ºC for 30 minutes. The mixture was then transferred to ice and an equal volume of PEG/NaCl (15% PEG, 1.6M NaCl) was added. This mixture was then incubated on ice for 5 minutes before centrifiguation at 14,000rpm for 5 minutes.

The supernatant was discarded and the pellet was resuspended in 300µl TE. The plasmid DNA was then purified by PhOH/$CHCl_3$ extraction. The DNA was precipitated by addition of 0.05 (v/v) 3M NaOAc (pH 5.2-5.6) and 2 (v/v) 100% ethanol. This was thoroughly mixed and incubated at -20ºC overnight. The tubes were then centrifuged at 14,000rpm for 5 minutes. The pellets were then washed with 70% ethanol, air-dried and resuspended in 300µl ddH_2O.

5.2.2.4 Removal of protein from DNA using Phenol/chloroform extraction

Water saturated distilled phenol (Rathburn chemicals, containing 0.1% hydroxyquinolone, was mixed with an equal volume of 0.5M Tris.Cl pH8 containing 0.2% mercaptoethanol. Prior to use, equilibrated phenol was mixed with an equal volume of chloroform. DNA to be extracted was added to an equal volume of this phenol/chloroform mixture and mixed thoroughly. The phases were separated by centrifugation in a microcentrifuge for 5 minutes. The aqueous phase was removed and extracted with an equal volume of chloroform to remove any residual phenol. After separation of the two resulting phases by centrifugation the DNA solution was removed to a new tube.

5.2.2.5 Precipitation of DNA using ethanol

DNA in solution was precipitated by the addition of 1/9 volume 3M sodium acetate pH 5.2 followed by 3 volumes of ethanol. After mixing, the solution was incubated for 20 minutes on ice and DNA recovered by centrifugation at 13,000rpm for 10 minutes in a Biofuge 13 microcentrifuge (Heraeus). Following removal of the supernatant the pellet was washed with 70% ethanol and dried for 10 minutes at room temperature. DNA was dissolved in a desirable volume of dH_2O or TE.

5.2.2.6 Quantification of DNA

DNA concentrations were estimated by measurement of absorption at 260nm using a lambda UV/VIS spectrophotometer (Perkin Elmer). Absorption measurements were converted to DNA concentrations using an extinction coefficient of 50µg/ml for double-stranded DNA and 33µg/ml for single-stranded DNA.

5.2.2.7 Cleavage of DNA by restriction endonucleases

DNA cleavage was carried out using enzymes and buffers supplied by Boehringer Mannheim and New England Biolabs under the conditions recommended by the manufacturers. Digests of 0.1 to 20µg DNA were carried out in 20-100µl of the appropriate reaction buffer for 1-12 hours at 37ºC.

5.2.2.8 Agarose gel electrophoresis

Electrophoresis of DNA was carried out in 0.7-2% MP (depending on the sizes of DNA fragments that were separated) agarose (Boehringer Mannheim) in TAE containing 0.5mg/ml ethidium bromide. Prior to loading, DNA samples were mixed with 1/6 volume 6_xagarose gel loading buffer. A potential difference of 1-10V per cm gel was used to separate DNA fragments. Following electrophoresis DNA was visualised and photographed on a UV transiluminator.

5.2.2.9 Purification of DNA fragments from agarose

Gel slices containing DNA fragments separated by agarose gel electrophoresis were purified using the QIAquick gel extraction kit (Qiagen) according to the manufacturer's instructions. This involves binding of DNA to a silica gel membrane at low pH in the presence of chaotropic salt, followed by washing in a buffer containing ethanol and low salt elution in 10mM Tris.CL pH8.5.

5.2.2.10 Ligation of DNA fragments 1

Ligation of PCR products into pGEM®-T was carried out according to the instructions of the manufacturer of the pGEM®-T vector system cloning kit (Promega).

5.2.2.11 Ligation of DNA fragments 2

In order to maximise the ligation between vector and insert fragments, a standard formula was used to predict the best fragment vector ratios.

[vector (ng) x fragment size (bp)/ vector size (bp)] x 3 = ng of insert needed

T4 DNA ligase (NEB) was used according to the manufacturer's instructions. Ligations were performed at 16ºC overnight.

5.2.2.12 Sequencing of double-stranded plasmid DNA

For sequencing reactions 500ng of DNA and 3.2pmole/µl were added to 3.68µl ddH_2O, 2µl 5_xSequencing Buffer and 2µl Big Dye and were subjected to 25 cycles of 95ºC for 30 seconds, 50ºC for 20 seconds and 60ºC for 4 minutes.

The samples were then cleaned up using Edge Biosystems Performa DTR V3 plates which remove dNTP's, salts, label from probes and other low molecular weight material. The samples were transferred to plates and dried down in a vacuum concentrator. 10µl of Hi-Di Formamide was added to each well and the plate put on a brief heat cycle of 95ºC for 2 minutes and cooled back to 4ºC. The plate was then put on the sequencer machine (3730 DNA Analyzer) and the samples were run on a 50cm array with POP-7 polymer. Thse reactions were performed by the Ashworth Sequencing Service in King's Buildings, Edinburgh.

5.2.2.13 Polymerase chain reaction

PCRs were carried out in 50µl of the appropriate PCR buffer with 0.5µM of each primer, 0.25mM of each dNTP (Boehringer Mannheim), and 1 unit of *Taq* DNA polymerase (Promega). Reactions were incubated at 95ºC for 10 minutes followed by 30 cycles of 1min at 95ºC, 1 minute at the annealing temperature, (generally 55ºC unless otherwise specified) and 1minute at 72ºC. Finally reactions were incubated at 72ºC for 10 minutes.

5.2.2.14 PCR product processing

PCR products were cloned using the pGEM®-T vector system I kit (Promega). This utilises a pre-cut plasmid vector (pGEM®-T) having a single unpaired deoxythymidine nucleotide at each 3' end. This provides compatible overhangs for ligation to PCR products as thermostable polymersases add an unpaired deoxyadenosine to the 5' end during synthesis. Following ligation and transformation, colonies with plasmids containing insertions were detected by blue-white selection.

5.2.2.15 Radiolabelling of oligonucleotides

Double stranded oligonucleotides were radioactively labelled with A-^{32}P-dCTP (3000Ci/mM) using Stratagene NUCtrap push-columns. Activity of oligonucleotides was measured through Cerenkoff counting.

5.2.2.16 Gel mobility shift assay for DNA–protein interactions

DNA mobility shift assays were performed asfollows. Unless otherwise indicated, cell extracts (30 μg) were incubated with radiolabelled oligonucleotides(2–3_x 10^5 c.p.m.) in 40 μl of binding buffer (pH 8.0), containing 5 μg poly dI-dC double stranded carrier DNA, 10 mM KCl, 10% glycerol and 1 mM DTT, for 10 min at 30°C before running on a polyacrylamide gel. For competition studies, unlabelled competitor DNA was added 15 min before radiolabelled DNA to protein fraction in the buffer described above. Following a 30 min incubation 30°C, 4 μl of 6_x native gel loading buffer (30% glycerol, 0.025% bromophenol blue and 0.025% xylene cyanol) were added, and the RNA–protein complexes were resolved on a 8% polyacrylamide gel. Gels were dried and exposed to X-ray film for 10–15 h with an intensifying screen at -70°C. All the experiments were performed at least three times.

5.2.3 Manipulation of *Drosophila melanogaster* flies and tissues

5.2.3.1 Fly stocks

Wild-type flies were of the Oregon R stock. For the Six4-III-GFP, Six4-3int-1/2GFP and Six4-3int-2/2GFP enhancer constructs, the primers TCTAGACAGCAAAGACCGTGAGTTG and GGATCCGAATGGATTGCCATCCAGTTG were used to amplify the *Six4* third intron sequence from wild-type genomic DNA, and the fragment was inserted into the pH Stinger vector (Barolo et al., 2000). The resulting plasmid was used to transform the w^{1118} strain by standard methods.

5.2.3.2 Maintenance of *Drosophila* stocks

Drosophila melanogaster strains were maintained at 25ºC on Dundee fly food. To maintain the reactivity of stocks, only flies up to seven days old were used for breeding.

5.2.3.3 Collection of *Drosophila* developmental stages

Embryos were collected on egg collection medium in Petri dishes placed on the bottom of fly cages. Plates were spread with yeast to provide food for the flies. After allowing females to lay eggs for the appropriate length of time, plates were collected and embryos were washed off onto nylon mesh with distilled water. Embryos were

then thoroughly washed with distilled water, and collected in Eppendorf tubes. Adult flies were anaesthetised (with carbon dioxide), sexed and collected in Eppendorf tubes.

5.2.3.4 Fixation of embryos for immunohistochemistry

Embryos were collected on grape juice plates with a globule of yeast paste (20% glucose) as a nutrient source. The grape juice plates were then aged for the appropriate length of time at appropriate temperatures (see table below). The embryos were removed using ddH2O and a paintbrush, and pipetted into a fine sieve. Embryos were washed to remove yeast and dechorionated in 50% fresh bleach for 4 minutes, they were then thoroughly washed to remove bleach. The embryos were then transferred into a scintillation vial and fixed for 20 minutes with agitation in 1.25ml formaldehyde (37%), 3.75ml PBS (8g NaCl, 0.2g KCLl, 1.44g Na2HPO4, 0.24g KH2PO4 for 1 litre, adjusted to pH 7.4) and 5 ml n-Heptane (Sigma). The bottom phase of formaldehyde was removed and 10ml of methanol was added. The scintillation vial was then shaken for 30 seconds to devitellinise the embryos. Embryos were allowed to settle to the bottom of the vial and then transferred to an Eppendorf microfuge tube. The embryos were then washed with methanol to remove residual heptane, and then washed 4 times with PBST. This was followed by the standard wash procedure.

Embryos were blocked for at least two hours in 2% bovine serum albumin (BSA) solution (Sigma) in PBST at room temperature on a rotating wheel. Primary antibody, in PBST at the appropriate concentration with 0.5% (v/v) BSA, 0.05% (v/v) Normal Goat Serum (NGS, Jackson labs) was added and samples were incubated at 4ºC overnight. The primary antibodies were then rinsed with the standard wash procedure. The secondary antibody (fluorochrome conjugate) was added in PBST to a concentration of 1:1000 for 2 hours at room temperature. The samples were rinsed with the standard wash procedure, they were then mounted in Vectrashield (Vector labs) on microscope slides sealed with a cover slip and nail varnish. Slides were stored in the dark at 4ºC. Confocal images were taken on a Zeiss LSM5 Pascal confocal microscope.

5.2.3.5 Preparation of *Drosophila* genomic DNA

20 flies were frozen for 5 minutes at -70ºC then resuspended in 400µl lysis buffer. The flies were homogenised using a hand-held Pellet-pestle® motor homogeniser (Kontes). Following incubation at 70ºC for 30 minutes, 56µl 8M potassium acetate was added. Samples were incubated 30 minutes on ice. To remove insoluble material, samples were centrifuged at 4ºC for 15 minutes at full speed in a microcentrifuge. The supernatant was removed and the centrifugation repeated. The final supernatant was added to 200µl of isopropanol and cooled at -70ºC for 10 minutes for precipitation of DNA. DNA was recovered by centrifugation, washed with 70%, dried and resuspended in 40µl TE.

5.2.3.6 Generation of transformant fly lines by microinjection

Constructs containing the SIX gene of interest in a pUASt vector (P element vector) were injected into Δ2-3 flies. The Δ 2-3 is the source of transposase for the attenuated P element vector. DNA is introduced into pre-cellular blastoderm embryos by injection and integrated into the genome by random transposition events. DNA for each construct was prepared using the method described above.

Cages of flies were set up and the grape-juice agar plated with yeast paste changed regularly to encourage egg laying. Plates were collected every hour and the embryos were used for injection. The injection procedure was carried out at 18ºC. Embryos were dechorionated for 4 minutes in 50% bleach and then rinsed in H_2O. Embryos were lined up under a microscope along the edge of a piece of agar in one orientation. They were then transferred to a coverslip coated with a film of glue. The coverslip was attached to a microscope slide using a drop of oil and placed at 18ºC for 20 minutes. It was then transferred to silica beads at 18ºC for 10 minutes to allow for dehydration. Embryos were then covered with series 700 halocarbon oil and injected with the construct of interest. Injected embryos were then covered in series 95 halocarbon oil, left at 18ºC for two days and then allowed to develop at 25ºC. Adult flies were crossed with white eyed flies (w1118) and transformants screened for on the basis of eye colour.

5.2.4 Immunohistochemistry

Antibody staining of whole-mount embryos was performed using standard methods and detected using secondary antibodies conjugated to Alexa 488 or 568 fluorochromes (Molecular Probes). Primary antibodies used Eya (mouse 1/100, Developmental Studies Hybridoma Bank, developed by N. Bonini) and GFP (mouse, 1/1000, Molecular Probes). For double labelling, RNA in situ hybridization was performed first followed by immunofluorescence.

5.2.5 Microscopy

Fluorescently labeed embryos were visualized by laser scanning confocal microscopy on either an SP (Leica) or a Pascal (Zeiss) microscope system. Images were processed and arranged using Adobe Photoshop software.

5.2.6 SELEX

5.2.6.1 GST-Six4 Recombinant Protein Expression and Purification

For pull-down assays, the SIX domain and homeodomain of Six4 was fused to glutathione S-transferase (GST) by cloning the respective amplified nucleotide sequences of Six4 into the pGEX-2T vector. Plasmids encoding GST plus the SIX and homeodomains of Six4 were transformed in *E.coli* strain BL21. Cells were grown overnight to a final OD of 1.3. Expression of fusion proteins was induced by adding isopropyl-β-D-thiogalactopyranoside to a final concentration of 0.05 mM for 3.5 h. The bacteria were harvested by centrifugation and resuspended in 5 ml of NTEN buffer (20 mM Tris pH 8, 100 mM NaCl, 1 mM EDTA, and 0.5% NP40) at 4 °C for 20 min. After sonication and centrifugation to remove cell debris, the supernatant was incubated with 200 µl of glutathione–sepharose beads (BD Bioscience, Franklin Lakes, NJ) at 4 °C for 1 h. After three washes in 5 ml of binding buffer (20 mM Tris pH 8, 100 mM KCl, 5 mM MgCl2, 0.1 mM EDTA, 20% glycerol, and 0.1% NP40), the levels of GST fusion proteins bound to the beads were checked by SDS–PAGE stained with Coomassie blue.

5.2.6.2 PCR amplification of selected oligonucleotides

Selected oligonucleotides were amplified using primers Primer1 and Primer2. Usually 1 unit of DyNAzyme EXT DNA Polymerase per 50 µl reaction volume gave good results, but for difficult templates (final SELEX rounds with large template concentrations) the optimal concentration was found to be 0.5 - 3 U per 50 µl reaction. These conditions were established after optimisation through the use of the Stratagene Opti-Prime^{TM} kit.

5.2.6.3 SELEX

DNA selection experiments were performed essentially as described (Treisman et al., 1991). An oligonucleotide library harbouring a variable random sequence surrounded by primer binding sites was rendered double stranded through PCR amplification and applied to a pre-column containing GST and glutathione–agarose in DNA-binding buffer (100pM)[25 mM Tris–HCl pH7.5, 200 mM KCl, 1 mM DTT, 0.05%, 40 µg/ml BSA], poly dI-dC double stranded carrier DNA (Amersham) was added to a final concentration of 20 µg/ml to absorb non-specifically-bound DNAs. Following washing in 3x 1ml of DNA-binding buffer, DNA was eluted in DNA-binding buffer containing 2M NaCl phenol extracted, ethanol precipitated, PCR amplified (15 cycles). PCR products were then used as template for the next round of selection. After five selection rounds, PCR products were subcloned for sequencing in the TOPO cloning vector.

5.2.6.4 Comparative quantification of recombinant protein yield

Coomassie protein assay reagent was obtained from Pierce. Soluble purified protein extracts were mixed with Coomassie reagent 1/31 to a final volume of 1.55 ml and incubated at room temperature for 5 minutes with moderate shaking at regular intervals. A spectrophotometer was blanked using a plastic cuvette filled with distilled water at a wavelength of 595 nm. Measurements of the absorbance were then taken for all the samples obtained under the different induction conditions. These measurements were

5.3 Statistical analyses

Statistical analyses were performed using the one-way analysis of variance followed by Fisher's protected least significant difference test. Statistical significance was assessed at $P<0.05$.

5.4 Utilised Algorithms and Websites

Most of the URLs of accessed websites are also included in the main text for convenience

Sequence retrieval and Data mining

FLYBASE: http://flybase.bio.indiana.edu/

ENSEMBL: http://www.ensembl.org/index.html

UCSC Genome Browser: http://genome.ucsc.edu/

REDFly2.0: http://redfly.ccr.buffalo.edu/

ORegAnno: www.oreganno.org/

FlyReg : FlyReg

Multiple Sequence alignment

CLUSTALW: http://www.ebi.ac.uk/Tools/clustalw/index.html

T-COFFEE: http://www.ebi.ac.uk/t-coffee/

MUSCLE: http://www.ebi.ac.uk/muscle/

Motif Elicitation

MEME: http://meme.nbcr.net/meme/meme.html

WEEDER: http://159.149.109.9/modtools/

See also consensus and AlignAce available through BEST as well as MotifScanner integrated in TOUCAN 2

Motif-based hidden Markov models utilities

http://metameme.sdsc.edu/cgi-bin/submit-verify.cgi

Hidden Markov Model generation and sequence scanning utilities

HMMER2: http://hmmer.janelia.org/

MAPPER: http://mapper.ChIP.org/

Matrix scanning utilities

Matrix-scan, Genome wide patser: accessed through RSAT

MatInspector: accessed through the GENOMATIX portal

PATCH: accessed through the gene-regulation portal

Transcription Factor Binding site identification

TESS:http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=WELCOME

Numerous utilities at: http://www.gene-regulation.com/

MATInspector: at http://www.genomatix.de/

BEST: http://www.cs.uga.edu/~che/BEST/

MOTIFScanner v3.1.1: www.esat.kuleuven.be/~thijs/Work/MotifScanner.html

Enhancer Modelling

Frameworker: at http://www.genomatix.de/

ModuleSearcher: accessed through TOUCAN 2

Orthologous gene Prediction

InParanoid: at http://inparanoid.sbc.su.se/cgi-bin/index.cgi

Additional Utilities and Websites

GENOMATIX portal: http://www.genomatix.de/

Gene-regulation portal: http://www.gene-regulation.com/

NCBI Toolbox: http://www.ncbi.nlm.nih.gov/Tools/

ExPASy Proteomics tools: http://www.expasy.ch/tools/

RSAT http://rsat.ulb.ac.be/rsat/

TOUCAN2: http://homes.esat.kuleuven.be/~saerts/software/toucan.php

Chapter 6 – Discussion and Concluding Remarks

6.1 Conclusions

The elucidation of the Six4 DNA binding specificity, the generation of a comprehensive positional weight matrix and the creation of a list of likely Six4 targets while not sufficient by themselves in providing a definitive answer to the question posed by the involvement of Six4 in *Drosophila* development are nonetheless a useful advancement. They provide a stepping stone from which to further query the role of SIX4/5 subfamily TFs in gonadogenesis and development in general. This study has filtered the existing body of information sufficiently for further *in vivo* and *in vitro* tests to be conducted and has in places begun to unravel the complex interactions underpinning the actions of Six4.

The consensus binding sequence of Six4 was found to be highly similar to that of its murine homologue Six5. Similarly, the permissive positional occupancy permutations were highly similar to those of Six5, thus suggesting a high degree of conservation in the pathways regulated by Six4 and allowing for cross-specific inferences to be made.

However, previously reported Six5 binding sites located within enhancers of reported Six5 targets were not always conserved in the homologous *Drosophila* sequences hinting towards considerable regulatory plasticity in the related developmental pathways between the two species. Another factor that can account for this phenomenon is the possible translocation of TFBSs outside the searched space. This possibility can be addressed through CRM detection analyses that will pinpoint the regulatory elements controlling the expression of homologous proteins more accurately.

A comprehensive search of all the suspected (or validated) CRMs of all *Drosophila* genes (as well as phylogenetically footprinted genomic regions) has uncovered a number of matches to the Six4 PWM. Experimental validation of these targets is still pending although some of these matches show considerable interspecific conservation within the *Drosophila* genus. Furthermore, a gene ontology analysis of these genes has uncovered strong patterns within the gene list. Many of these genes have been shown to have functions that are highly consistent with Six4 activity. The statistical significance of these findings assigns additional confidence to any inferences about the identity of Six4 targets based on them.

Contrary to expectations the findings of this study seem to refute the claim of Six4 being under the direct regulation of the Zfh-1 transcription factor (Clark et al., 2006)

since no putative binding sequences for this factor were discovered within Six4-3int, the enhancer element responsible for the mesodermal expression of Six4. This suggests that any regulation of Six4 by this factor must be carried out through the action of mediating transcription factors.

On the contrary, a phylogenetically conserved putative TFBS of the mesodermal patterning transcription factor Tinman, was discovered in the Six4-3int enhancer. Additional evidence provided by an enhancer partitioning analysis suggests a role for Tinman in the regulation of *Six4*. This finding is in agreement with the proposed role for Tinman in the regulation of *Six4* expression (Clark et al., 2006) as well as more recent findings by others (Eileen Furlong, personal communication). My findings also hint towards the regulation of Six4 by the pan-mesodermal transcription factor Twist.

Additionally, the same enhancer partitioning analysis of Six4-3int has shown that the entirety of the Six4-3int enhancer is required to confer the complete *Six4* mesodermal expression pattern. Ablation of either half of the enhancer results in atypical expression of a GFP reporter gene with the majority of the gonadal expression pattern being retained by the 5' half of the Six4-3int enhancer. Additionally a similar yet distinct expression pattern (to that conferred by Six4-3int) is also associated with the ablation of the 5' end of the Six4-3int enhancer, thus hinting towards the presence of repressor binding sites in the 5' half of the enhancer responsible for suppressing this expression pattern.

The disruption of the Six4-3int expression pattern can potentially hint towards the functional utility of putative binding sites such as those of Tinman and Twist based on the relative position of those putative TFBSs within the enhancer. A more comprehensive enhancer partitioning or binding site mutagenesis analysis is however required to either validate or refute this claim. Ideally such an analysis would involve the disruption of just the putative TFBSs in question

Additionally, this study has constructed a list of genes, many of which currently have no assigned function, which are expressed in a number of patterns consistent with that of *Six4*. Initial putative TFBS analyses have failed to convincingly identify the factors responsible for this expression pattern, although this gene list is a useful starting point for a more in-depth common TFBS identification analysis.

Finally, this study makes use of the existing body of evidence as well as its own findings to place Six4 within a network of interactions that ultimately control gonadogenesis through processes up to, and including, SGP specification. Elucidation of this network will allow us to infer the functions of homologous factors in the

development of higher eukaryotes leading up to and including humans as well as to attain a clearer understanding of the pathology of disorders arising from the disruption of development such as myotonic dystrophy.

6.2 Future experiments

Validation of the putative regulatory sequences identified is the next logical step in furthering this study. The use of enhancer driven reporter genes in *Drosophila* transformants in conjunction with binding site deletion and/or mutagenesis can provide definitive evidence concerning the authenticity of these elements. This approach can be equally applied to potential Six4 target sequences and to putative TFBSs identified in the Six4-3int enhancer as well as the putative regulatory elements consisting of both Tinman and Twist binding sites.

Immunohistochemistry using a Six4 antibody in *Drosophila* strains mutant for the genes suspected of *Six4* regulation could substantiate or disprove the theories proposed by his study concerning *Six4* regulation. Alternatively, Six4-3int-GFP expressing lines can be crossed with strains mutant for the genes in question in an attempt to assay GFP expression.

Additionally, the *in silico* approaches utilised in this study to identify the common factor(s) responsible for conferring the *Six4*-like expression pattern are not exhaustive and could be expanded to incorporate information such as phylogenetic footprinting and/or TFBS positioning within putative CRMs.

These proposed experiments do not exhaust the available possibilities but they address the immediate concerns stemming from the findings of this study.

Bibliography

The Gene Ontology project in 2008. Nucleic Acids Res 36, D440-4 (2008).

Abel, T., Michelson, A. M. & Maniatis, T. A Drosophila GATA family member that binds to Adh regulatory sequences is expressed in the developing fat body. Development 119, 623-33 (1993).

Aerts, S., Van Loo, P., Moreau, Y. & De Moor, B. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. Bioinformatics 20, 1974-6 (2004).

Aerts, S., Van Loo, P., Thijs, G., Mayer, H., de Martin, R., Moreau, Y. & De Moor, B. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. Nucleic Acids Res 33, W393-6 (2005).

Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. & De Moor, B. Computational detection of cis -regulatory modules. Bioinformatics 19 Suppl 2, ii5-14 (2003).

Alkema, W. B., Johansson, O., Lagergren, J. & Wasserman, W. W. MSCAN: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res 32, W195-8 (2004).

Almon, R. R., Lai, W., DuBois, D. C. & Jusko, W. J. Corticosteroid-regulated genes in rat kidney: mining time series array data. Am J Physiol Endocrinol Metab 289, E870-82 (2005).

Alwazzan, M., Hamshere, M. G., Lennon, G. G. & Brook, J. D. Six transcripts map within 200 kilobases of the myotonic dystrophy expanded repeat. Mamm Genome 9, 485-7 (1998).

Andrews, P. I. & Wilson, J. Relative disease severity in siblings with myotonic dystrophy. J Child Neurol 7, 161-7 (1992).

Aoyagi, N. & Wassarman, D. A. Genes encoding Drosophila melanogaster RNA polymerase II general transcription factors: diversity in TFIIA and TFIID components contributes to gene-specific transcriptional regulation. J Cell Biol 150, F45-50 (2000).

Arnone, M. I. & Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. Development 124, 1851-64 (1997).

Arnosti, D. N. Design and function of transcriptional switches in Drosophila. Insect Biochem Mol Biol 32, 1257-73 (2002).

Arnosti, D. N. Analysis and function of transcriptional regulatory elements: insights from Drosophila. Annu Rev Entomol 48, 579-602 (2003).

Aslanidis, C., Jansen, G., Amemiya, C., Shutler, G., Mahadevan, M., Tsilfidis, C., Chen, C., Alleman, J., Wormskamp, N. G., Vooijs, M. & et al. Cloning of the essential myotonic dystrophy region and mapping of the putative defect. Nature 355, 548-51 (1992).

Ausubel, L. J., Kwan, C. K., Sette, A., Kuchroo, V. & Hafler, D. A. Complementary mutations in an antigenic peptide allow for crossreactivity of autoreactive T-cell clones. Proc Natl Acad Sci U S A 93, 15317-22 (1996).

Azpiazu, N. & Frasch, M. tinman and bagpipe: two homeo box genes that determine cell fates in the dorsal mesoderm of Drosophila. Genes Dev 7, 1325-40 (1993).

Azpiazu, N., Lawrence, P. A., Vincent, J. P. & Frasch, M. Segmentation and specification of the Drosophila mesoderm. Genes Dev 10, 3183-94 (1996).

Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2, 28-36 (1994).

Barbosa, J., Nuttall, F. Q., Kennedy, W. & Goetz, F. Plasma insulin in patients with myotonic dystrophy and their relatives. Medicine (Baltimore) 53, 307-23 (1974).

Barolo, S., Carver, L. A. & Posakony, J. W. GFP and beta-galactosidase transformation vectors for promoter/enhancer analysis in Drosophila. Biotechniques 29, 726, 728, 730, 732 (2000).

Bartel, D. P. & Szostak, J. W. Isolation of new ribozymes from a large pool of random sequences [see comment]. Science 261, 1411-8 (1993).

Bauer, S., Grossmann, S., Vingron, M. & Robinson, P. N. Ontologizer 2.0 - A Multifunctional Tool for GO Term Enrichment Analysis and Data Exploration. Bioinformatics (2008).

Baylies, M. K. & Bate, M. twist: a myogenic switch in Drosophila. Science 272, 1481-4 (1996).

Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S. & Grosse, I. Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics 21, 2657-66 (2005).

Benos, P. V., Corcoran, D. L. & Feingold, E. Web-based identification of evolutionary conserved DNA cis-regulatory elements. Methods Mol Biol 395, 425-36 (2007).

Benos, P. V., Lapedes, A. S., Fields, D. S. & Stormo, G. D. SAMIE: statistical algorithm for modeling interaction energies. Pac Symp Biocomput, 115-26 (2001).

Benson, M. & Pirrotta, V. The Drosophila zeste protein binds cooperatively to sites in many gene regulatory regions: implications for transvection and gene regulation. Embo J 7, 3907-15 (1988).

Berezikov, E., Guryev, V., Plasterk, R. H. & Cuppen, E. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. Genome Res 14, 170-8 (2004).

Berezovski, M., Musheev, M., Drabovich, A. & Krylov, S. N. Non-SELEX selection of aptamers. J Am Chem Soc 128, 1410-1 (2006).

Bergman, C. M., Carlson, J. W. & Celniker, S. E. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 21, 1747-9 (2005).

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M. & Eisen, M. B. 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 99, 757-62 (2002).

Berul, C. I., Maguire, C. T., Aronovitz, M. J., Greenwood, J., Miller, C., Gehrmann, J., Housman, D., Mendelsohn, M. E. & Reddy, S. DMPK dosage alterations result in atrioventricular conduction abnormalities in a mouse myotonic dystrophy model. J Clin Invest 103, R1-7 (1999).

Bibby, D. F., Gill, A. C., Kirby, L., Farquhar, C. F., Bruce, M. E. & Garson, J. A. Application of a novel in vitro selection technique to isolate and characterise high affinity DNA aptamers binding mammalian prion proteins. J Virol Methods (2008).

Billeter, M., Guntert, P., Luginbuhl, P. & Wuthrich, K. Hydration and DNA recognition by homeodomains. Cell 85, 1057-65 (1996).

Blackman, R. K., Sanicola, M., Raftery, L. A., Gillevet, T. & Gelbart, W. M. An extensive 3' cis-regulatory region directs the imaginal disk expression of decapentaplegic, a member of the TGF-beta family in Drosophila. Development 111, 657-66 (1991).

Blackwell, T. K., Kretzner, L., Blackwood, E. M., Eisenman, R. N. & Weintraub, H. Sequence-specific DNA binding by the c-Myc protein. Science 250, 1149-51 (1990).

Blanchette, M. & Tompa, M. FootPrinter: A program designed for phylogenetic footprinting. Nucleic Acids Res 31, 3840-2 (2003).

Bodmer, R. The gene tinman is required for specification of the heart and visceral muscles in Drosophila. Development 118, 719-29 (1993).

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L. & Rubin, E. M. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299, 1391-4 (2003).

Botquin, V., Hess, H., Fuhrmann, G., Anastassiadis, C., Gross, M. K., Vriend, G. & Scholer, H. R. New POU dimer configuration mediates antagonistic control of an osteopontin preimplantation enhancer by Oct-4 and Sox-2. Genes Dev 12, 2073-90 (1998).

Boucher, C. A., King, S. K., Carey, N., Krahe, R., Winchester, C. L., Rahman, S., Creavin, T., Meghji, P., Bailey, M. E., Chartier, F. L. & et al. A novel homeodomain-encoding gene is associated with a large CpG island interrupted by the myotonic dystrophy unstable (CTG)n repeat. Hum Mol Genet 4, 1919-25 (1995).

Boyle, M., Bonini, N. & DiNardo, S. Expression and function of clift in the development of somatic gonadal precursors within the Drosophila mesoderm. Development 124, 971-82 (1997).

Boyle, M. & DiNardo, S. Specification, migration and assembly of the somatic cells of the Drosophila gonad. Development 121, 1815-25 (1995).

Brody, T., Rasband, W., Baler, K., Kuzin, A., Kundu, M. & Odenwald, W. F. cis-Decoder discovers constellations of conserved DNA sequences shared among tissue-specific enhancers. Genome Biol 8, R75 (2007).

Broihier, H. T., Moore, L. A., Van Doren, M., Newman, S. & Lehmann, R. zfh-1 is required for germ cell migration and gonadal mesoderm development in Drosophila. Development 125, 655-66 (1998).

Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J. P., Hudson, T. & et al. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell 69, 385 (1992).

Brookman, J. J., Toosy, A. T., Shashidhara, L. S. & White, R. A. The 412 retrotransposon and the development of gonadal mesoderm in Drosophila. Development 116, 1185-92 (1992).

Bryne, J. C., Valen, E., Tang, M. H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B. & Sandelin, A. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36, D102-6 (2008).

Buchanan, C. D., Klein, P. E. & Mullet, J. E. Phylogenetic analysis of 5'-noncoding regions from the ABA-responsive rab16/17 gene family of sorghum, maize

and rice provides insight into the composition, organization and function of cis-regulatory modules. Genetics 168, 1639-54 (2004).

Bui, Q. T., Zimmerman, J. E., Liu, H., Gray-Board, G. L. & Bonini, N. M. Functional analysis of an eye enhancer of the Drosophila eyes absent gene: differential regulation by eye specification genes. Dev Biol 221, 355-64 (2000).

Bu'Lock, F. A., Sood, M., De Giovanni, J. V. & Green, S. H. Left ventricular diastolic function in congenital myotonic dystrophy. Arch Dis Child 80, 267-70 (1999).

Bulyk, M. L. Analysis of sequence specificities of DNA-binding proteins with protein binding microarrays. Methods Enzymol 410, 279-99 (2006).

Bulyk, M. L. DNA microarray technologies for measuring protein-DNA interactions. Curr Opin Biotechnol 17, 422-30 (2006).

Bulyk, M. L., Johnson, P. L. & Church, G. M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res 30, 1255-61 (2002).

Burtis, K. C., Coschigano, K. T., Baker, B. S. & Wensink, P. C. The doublesex proteins of Drosophila melanogaster bind directly to a sex-specific yolk protein gene enhancer. Embo J 10, 2577-82 (1991).

Busturia, A. & Bienz, M. Silencers in abdominal-B, a homeotic Drosophila gene. Embo J 12, 1415-25 (1993).

Camargo, A., Azuaje, F., Wang, H. & Zheng, H. Permutation - based statistical tests for multiple hypotheses. Source Code Biol Med 3, 15 (2008).

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. & Werner, T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21, 2933-42 (2005).

Cavener, D. R. & Ray, S. C. Eukaryotic start and stop translation sites. Nucleic Acids Res 19, 3185-92 (1991).

Censori, B., Provinciali, L., Danni, M., Chiaramoni, L., Maricotti, M., Foschi, N., Del Pesce, M. & Salvolini, U. Brain involvement in myotonic dystrophy: MRI features and their relationship to clinical and cognitive conditions. Acta Neurol Scand 90, 211-7 (1994).

Chakrabarti, S., Lanczycki, C. J., Panchenko, A. R., Przytycka, T. M., Thiessen, P. A. & Bryant, S. H. State of the art: refinement of multiple sequence alignments. BMC Bioinformatics 7, 499 (2006).

Cheng, T. C., Wallace, M. C., Merlie, J. P. & Olson, E. N. Separable regulatory elements governing myogenin transcription in mouse embryogenesis. Science 261, 215-8 (1993).

Cheyette, B. N., Green, P. J., Martin, K., Garren, H., Hartenstein, V. & Zipursky, S. L. The Drosophila sine oculis locus encodes a homeodomain-containing protein required for the development of the entire visual system. Neuron 12, 977-96 (1994).

Chi, Y. I. Homeodomain revisited: a lesson from disease-causing mutations. Hum Genet 116, 433-44 (2005).

Choi, Y. S. & Sinha, S. Determination of the consensus DNA-binding sequence and a transcriptional activation domain for ESE-2. Biochem J 398, 497-507 (2006).

Clark, I. B., Boyd, J., Hamilton, G., Finnegan, D. J. & Jarman, A. P. D-six4 plays a key role in patterning cell identities deriving from the Drosophila mesoderm. Dev Biol 294, 220-31 (2006).

Clark, I. B., Jarman, A. P. & Finnegan, D. J. Live imaging of Drosophila gonad formation reveals roles for Six4 in regulating germline and somatic cell migration. BMC Dev Biol 7, 52 (2007).

Cobo, A., Martinez, J. M., Martorell, L., Baiget, M. & Johnson, K. Molecular diagnosis of homozygous myotonic dystrophy in two asymptomatic sisters. Hum Mol Genet 2, 711-5 (1993).

Collas, P. & Dahl, J. A. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. Front Biosci 13, 929-43 (2008).

Corcoran, D. L., Feingold, E. & Benos, P. V. FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. Nucleic Acids Res 33, W442-6 (2005).

Costas, J., Valade, E. & Naveira, H. Structural features of the mdg1 lineage of the Ty3/gypsy group of LTR retrotransposons inferred from the phylogenetic analyses of its open reading frames. J Mol Evol 53, 165-71 (2001).

Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res 14, 1188-90 (2004).

Cumberledge, S., Szabad, J. & Sakonju, S. Gonad formation and development requires the abd-A domain of the bithorax complex in Drosophila melanogaster. Development 115, 395-402 (1992).

Cuticchia, A. J., Ivarie, R. & Arnold, J. The application of Markov chain analysis to oligonucleotide frequency prediction and physical mapping of Drosophila melanogaster. Nucleic Acids Res 20, 3651-7 (1992).

Damiani, E., Angelini, C., Pelosi, M., Sacchetto, R., Bortoloso, E. & Margreth, A. Skeletal muscle sarcoplasmic reticulum phenotype in myotonic dystrophy. Neuromuscul Disord 6, 33-47 (1996).

Das, M. K. & Dai, H. K. A survey of DNA motif finding algorithms. BMC Bioinformatics 8 Suppl 7, S21 (2007).

Dasgupta, P. & Chellappan, S. P. Chromatin immunoprecipitation assays: molecular analysis of chromatin modification and gene regulation. Methods Mol Biol 383, 135-52 (2007).

David, R., Ahrens, K., Wedlich, D. & Schlosser, G. Xenopus Eya1 demarcates all neurogenic placodes as well as migrating hypaxial muscle precursors. Mech Dev 103, 189-92 (2001).

Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. & Lempicki, R. A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4, P3 (2003).

Di Costanzo, A., Santoro, L., de Cristofaro, M., Manganelli, F., Di Salle, F. & Tedeschi, G. Familial aggregation of white matter lesions in myotonic dystrophy type 1. Neuromuscul Disord 18, 299-305 (2008).

Djordjevic, M. SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways. Biomol Eng (2007).

Djordjevic, M. & Sengupta, A. M. Quantitative modeling and data analysis of SELEX experiments. Phys Biol 3, 13-28 (2006).

Duboule, D. & Morata, G. Colinearity and functional hierarchy among genes of the homeotic complexes. Trends Genet 10, 358-64 (1994).

Eddy, S. R. Profile hidden Markov models. Bioinformatics 14, 755-63 (1998).

Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. Nature 346, 818-22 (1990).

Elnitski, L., Jin, V. X., Farnham, P. J. & Jones, S. J. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16, 1455-64 (2006).

Fan, X., Brass, L. F., Poncz, M., Spitz, F., Maire, P. & Manning, D. R. The alpha subunits of Gz and Gi interact with the eyes absent transcription cofactor Eya2,

preventing its interaction with the six class of homeodomain-containing proteins. J Biol Chem 275, 32129-34 (2000).

Fickett, J. W. & Wasserman, W. W. Discovery and modeling of transcriptional regulatory regions. Curr Opin Biotechnol 11, 19-24 (2000).

Filippova, G. N., Thienes, C. P., Penn, B. H., Cho, D. H., Hu, Y. J., Moore, J. M., Klesert, T. R., Lobanenkov, V. V. & Tapscott, S. J. CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. Nat Genet 28, 335-43 (2001).

Fogel, G. B., Weekes, D. G., Varga, G., Dow, E. R., Craven, A. M., Harlow, H. B., Su, E. W., Onyia, J. E. & Su, C. A statistical analysis of the TRANSFAC database. Biosystems 81, 137-54 (2005).

Foiry, L., Dong, L., Savouret, C., Hubert, L., te Riele, H., Junien, C. & Gourdon, G. Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. Hum Genet 119, 520-6 (2006).

Fougerousse, F., Durand, M., Lopez, S., Suel, L., Demignon, J., Thornton, C., Ozaki, H., Kawakami, K., Barbet, P., Beckmann, J. S. & Maire, P. Six and Eya expression during human somitogenesis and MyoD gene family activation. J Muscle Res Cell Motil 23, 255-64 (2002).

Frasch, M. Induction of visceral and cardiac mesoderm by ectodermal Dpp in the early Drosophila embryo. Nature 374, 464-7 (1995).

Friedrich, B., Quensel, C., Sommer, T., Hartmann, E. & Kohler, M. Nuclear localization signal and protein context both mediate importin alpha specificity of nuclear import substrates. Mol Cell Biol 26, 8697-709 (2006).

Frith, M. C., Hansen, U., Spouge, J. L. & Weng, Z. Finding functional sequence elements by multiple local alignment. Nucleic Acids Res 32, 189-200 (2004).

Fu, Y. & Weng, Z. Improvement of TRANSFAC matrices using multiple local alignment of transcription factor binding site sequences. Conf Proc IEEE Eng Med Biol Soc 4, 2856-9 (2004).

Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5, 3157-70 (1978).

Garabedian, M. J., Shepherd, B. M. & Wensink, P. C. A tissue-specific transcription enhancer from the Drosophila yolk protein 1 gene. Cell 45, 859-67 (1986).

Garcia-Fernandez, J. The genesis and evolution of homeobox gene clusters. Nat Rev Genet 6, 881-92 (2005).

Garcia-Lopez, A., Monferrer, L., Garcia-Alcover, I., Vicente-Crespo, M., Alvarez-Abril, M. C. & Artero, R. D. Genetic and chemical modifiers of a CUG toxicity model in Drosophila. PLoS ONE 3, e1595 (2008).

Gehring, W. [Cell heredity and changes of determination in cultures of imaginal discs in Drosophila melanogaster]. J Embryol Exp Morphol 15, 77-111 (1966).

Gehring, W. J., Affolter, M. & Burglin, T. Homeodomain proteins. Annu Rev Biochem 63, 487-526 (1994).

Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resendez-Perez, D., Affolter, M., Otting, G. & Wuthrich, K. Homeodomain-DNA recognition. Cell 78, 211-23 (1994).

Giordani, J., Bajard, L., Demignon, J., Daubas, P., Buckingham, M. & Maire, P. Six proteins regulate the activation of Myf5 expression in embryonic mouse limbs. Proc Natl Acad Sci U S A (2007).

Gold, L. The SELEX process: a surprising source of therapeutic and diagnostic compounds. Harvey Lect 91, 47-57 (1995).

Gomes-Pereira, M., Foiry, L., Nicole, A., Huguet, A., Junien, C., Munnich, A. & Gourdon, G. CTG trinucleotide repeat "big jumps": large expansions, small mice. PLoS Genet 3, e52 (2007).

Gopinath, S. C. Methods developed for SELEX. Anal Bioanal Chem 387, 171-82 (2007).

Grad, Y. H., Roth, F. P., Halfon, M. S. & Church, G. M. Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura. Bioinformatics 20, 2738-50 (2004).

Gribskov, M. & Veretnik, S. Identification of sequence pattern with profile analysis. Methods Enzymol 266, 198-212 (1996).

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M. C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S. M., Giardine, B., Hooghe, B., Van Loo, P., Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E., Donaldson, I. J., Robertson, G., Wadelius, C., De Bleser, P., Vlieghe, D., Halfon, M. S., Wasserman, W., Hardison, R., Bergman, C. M. & Jones, S. J. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36, D107-13 (2008).

Grifone, R., Demignon, J., Houbron, C., Souil, E., Niro, C., Seller, M. J., Hamard, G. & Maire, P. Six1 and Six4 homeoproteins are required for Pax3 and Mrf

expression during myogenesis in the mouse embryo. Development 132, 2235-49 (2005).

Grifone, R., Laclef, C., Spitz, F., Lopez, S., Demignon, J., Guidotti, J. E., Kawakami, K., Xu, P. X., Kelly, R., Petrof, B. J., Daegelen, D., Concordet, J. P. & Maire, P. Six1 and Eya1 expression can reprogram adult muscle from the slow-twitch phenotype into the fast-twitch phenotype. Mol Cell Biol 24, 6253-67 (2004).

Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics 23, 3024-31 (2007).

Gutjahr, T., Vanario-Alonso, C. E., Pick, L. & Noll, M. Multiple regulatory elements direct the complex expression pattern of the Drosophila segmentation gene paired. Mech Dev 48, 119-28 (1994).

Halfon, M. S., Gallo, S. M. & Bergman, C. M. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. Nucleic Acids Res 36, D594-8 (2008).

Hampshire, A. J., Rusling, D. A., Broughton-Head, V. J. & Fox, K. R. Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. Methods 42, 128-40 (2007).

Hanahan, D. Studies on transformation of Escherichia coli with plasmids. J Mol Biol 166, 557-80 (1983).

Hannenhalli, S. Eukaryotic transcription factor binding sites--modeling and integrative search methods. Bioinformatics 24, 1325-31 (2008).

Hanson, I. M. Mammalian homologues of the Drosophila eye specification genes. Semin Cell Dev Biol 12, 475-84 (2001).

Harada, R., Dufort, D., Denis-Larose, C. & Nepveu, A. Conserved cut repeats in the human cut homeodomain protein function as DNA binding domains. J Biol Chem 269, 2062-7 (1994).

Harley, H. G., Brook, J. D., Rundle, S. A., Crow, S., Reardon, W., Buckler, A. J., Harper, P. S., Housman, D. E. & Shaw, D. J. Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. Nature 355, 545-6 (1992).

Harper, P. S. Myotonic dystrophy (ed. Tapscott, S. J.) (W.B. Saunders, London, 2001).

Harper, P. S. & Dyken, P. R. Early-onset dystrophia myotonica. Evidence supporting a maternal environmental factor. Lancet 2, 53-5 (1972).

Harris, S. E., Winchester, C. L. & Johnson, K. J. Functional analysis of the homeodomain protein SIX5. Nucleic Acids Res 28, 1871-8 (2000).

Heanue, T. A., Reshef, R., Davis, R. J., Mardon, G., Oliver, G., Tomarev, S., Lassar, A. B. & Tabin, C. J. Synergistic regulation of vertebrate muscle development by Dach2, Eya2, and Six1, homologs of genes required for Drosophila eye formation. Genes Dev 13, 3231-43 (1999).

Heery, D. M., Powell, R., Gannon, F. & Dunican, L. K. Curing of a plasmid from E.coli using high-voltage electroporation. Nucleic Acids Res 17, 10131 (1989).

Henikoff, J. G. & Henikoff, S. Using substitution probabilities to improve position-specific scoring matrices. Comput Appl Biosci 12, 135-43 (1996).

Himeda, C. L., Ranish, J. A., Angello, J. C., Maire, P., Aebersold, R. & Hauschka, S. D. Quantitative proteomic identification of six4 as the trex-binding factor in the muscle creatine kinase enhancer. Mol Cell Biol 24, 2132-43 (2004).

Himeda, C. L., Ranish, J. A. & Hauschka, S. D. Quantitative proteomic identification of MAZ as a transcriptional regulator of muscle-specific genes in skeletal and cardiac myocytes. Mol Cell Biol (2008).

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., Sultan-Qurraie, A., Thomas, D. J., Trumbower, H., Weber, R. J., Weirauch, M., Zweig, A. S., Haussler, D. & Kent, W. J. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34, D590-8 (2006).

Hobert, O. & Westphal, H. Functions of LIM-homeobox genes. Trends Genet 16, 75-83 (2000).

Hoskins, B. E., Cramer, C. H., Silvius, D., Zou, D., Raymond, R. M., Orten, D. J., Kimberling, W. J., Smith, R. J., Weil, D., Petit, C., Otto, E. A., Xu, P. X. & Hildebrandt, F. Transcription factor SIX5 is mutated in patients with branchio-oto-renal syndrome. Am J Hum Genet 80, 800-4 (2007).

Hsiao, F. C., Williams, A., Davies, E. L. & Rebay, I. Eyes absent mediates cross-talk between retinal determination genes and the receptor tyrosine kinase signaling pathway. Dev Cell 1, 51-61 (2001).

Hu, S., Mamedova, A. & Hegde, R. S. DNA-binding and regulation mechanisms of the SIX family of retinal determination proteins. Biochemistry 47, 3586-94 (2008).

Huang, H., Kao, M. C., Zhou, X., Liu, J. S. & Wong, W. H. Determination of local statistical significance of patterns in Markov sequences with application to promoter element identification. J Comput Biol 11, 1-14 (2004).

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Pagni, M. & Sigrist, C. J. The PROSITE database. Nucleic Acids Res 34, D227-30 (2006).

Impey, S., McCorkle, S. R., Cha-Molstad, H., Dwyer, J. M., Yochum, G. S., Boss, J. M., McWeeney, S., Dunn, J. J., Mandel, G. & Goodman, R. H. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. Cell 119, 1041-54 (2004).

Innis, M. A., Myambo, K. B., Gelfand, D. H. & Brow, M. A. DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. Proc Natl Acad Sci U S A 85, 9436-40 (1988).

Inoue, H., Nojima, H. & Okayama, H. High efficiency transformation of Escherichia coli with plasmids. Gene 96, 23-8 (1990).

Irvine, D., Tuerk, C. & Gold, L. SELEXION. Systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis. J Mol Biol 222, 739-61 (1991).

Istrail, S. & Davidson, E. H. Logic functions of the genomic cis-regulatory code. Proc Natl Acad Sci U S A 102, 4954-9 (2005).

Jamal, G. A., Weir, A. I., Hansen, S. & Ballantyne, J. P. Myotonic dystrophy. A reassessment by conventional and more recently introduced neurophysiological techniques. Brain 109 (Pt 6), 1279-96 (1986).

Janky, R. & van Helden, J. Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. BMC Bioinformatics 9, 37 (2008).

Jansen, G., Bartolomei, M., Kalscheuer, V., Merkx, G., Wormskamp, N., Mariman, E., Smeets, D., Ropers, H. H. & Wieringa, B. No imprinting involved in the expression of DM-kinase mRNAs in mouse and human tissues. Hum Mol Genet 2, 1221-7 (1993).

Jean, D., Bernier, G. & Gruss, P. Six6 (Optx2) is a novel murine Six3-related homeobox gene that demarcates the presumptive pituitary/hypothalamic axis and the ventral optic stalk. Mech Dev 84, 31-40 (1999).

Jemc, J. & Rebay, I. The eyes absent family of phosphotyrosine phosphatases: properties and roles in developmental regulation of transcription. Annu Rev Biochem 76, 513-38 (2007).

Johansson, O., Alkema, W., Wasserman, W. W. & Lagergren, J. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. Bioinformatics 19 Suppl 1, i169-76 (2003).

Kadonaga, J. T. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116, 247-57 (2004).

Kaelin, W. G., Jr., Krek, W., Sellers, W. R., DeCaprio, J. A., Ajchenbaum, F., Fuchs, C. S., Chittenden, T., Li, Y., Farnham, P. J., Blanar, M. A. & et al. Expression cloning of a cDNA encoding a retinoblastoma-binding protein with E2F-like properties. Cell 70, 351-64 (1992).

Kawakami, K., Ohto, H., Ikeda, K. & Roeder, R. G. Structure, function and expression of a murine homeobox protein AREC3, a homologue of Drosophila sine oculis gene product, and implication in development. Nucleic Acids Res 24, 303-10 (1996).

Kawakami, K., Sato, S., Ozaki, H. & Ikeda, K. Six family genes--structure and function as transcription factors and their roles in development. Bioessays 22, 616-26 (2000).

Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V. & Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31, 3576-9 (2003).

Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).

Khatri, P. & Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21, 3587-95 (2005).

Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res 17, 1919-31 (2007).

Kirby, R. J., Hamilton, G. M., Finnegan, D. J., Johnson, K. J. & Jarman, A. P. Drosophila homolog of the myotonic dystrophy-associated gene, SIX5, is required for muscle and gonad development. Curr Biol 11, 1044-9 (2001).

Klepper, K., Sandve, G. K., Abul, O., Johansen, J. & Drablos, F. Assessment of composite motif discovery methods. BMC Bioinformatics 9, 123 (2008).

Klesert, T. R., Cho, D. H., Clark, J. I., Maylie, J., Adelman, J., Snider, L., Yuen, E. C., Soriano, P. & Tapscott, S. J. Mice deficient in Six5 develop cataracts: implications for myotonic dystrophy. Nat Genet 25, 105-9 (2000).

Klesert, T. R., Otten, A. D., Bird, T. D. & Tapscott, S. J. Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of DMAHP. Nat Genet 16, 402-6 (1997).

Klug, S. J. & Famulok, M. All you wanted to know about SELEX. Mol Biol Rep 20, 97-107 (1994).

Kobayashi, M., Toyama, R., Takeda, H., Dawid, I. B. & Kawakami, K. Overexpression of the forebrain-specific homeobox gene six3 induces rostral forebrain enlargement in zebrafish. Development 125, 2973-82 (1998).

Kokoza, V. A., Martin, D., Mienaltowski, M. J., Ahmed, A., Morton, C. M. & Raikhel, A. S. Transcriptional regulation of the mosquito vitellogenin gene via a blood meal-triggered cascade. Gene 274, 47-65 (2001).

Korade-Mirnics, Z., Tarleton, J., Servidei, S., Casey, R. R., Gennarelli, M., Pegoraro, E., Angelini, C. & Hoffman, E. P. Myotonic dystrophy: tissue-specific effect of somatic CTG expansions on allele-specific DMAHP/SIX5 expression. Hum Mol Genet 8, 1017-23 (1999).

Kozmik, Z., Holland, N. D., Kreslova, J., Oliveri, D., Schubert, M., Jonasova, K., Holland, L. Z., Pestarino, M., Benes, V. & Candiani, S. Pax-Six-Eya-Dach network during amphioxus development: conservation in vitro but context specificity in vivo. Dev Biol 306, 143-59 (2007).

Kuhn, R. M., Karolchik, D., Zweig, A. S., Trumbower, H., Thomas, D. J., Thakkapallayil, A., Sugnet, C. W., Stanke, M., Smith, K. E., Siepel, A., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pedersen, J. S., Hsu, F., Hinrichs, A. S., Harte, R. A., Diekhans, M., Clawson, H., Bejerano, G., Barber, G. P., Baertsch, R., Haussler, D. & Kent, W. J. The UCSC genome browser database: update 2007. Nucleic Acids Res 35, D668-73 (2007).

Laclef, C., Hamard, G., Demignon, J., Souil, E., Houbron, C. & Maire, P. Altered myogenesis in Six1-deficient mice. Development 130, 2239-52 (2003).

Lagutin, O. V., Zhu, C. C., Kobayashi, D., Topczewski, J., Shimamura, K., Puelles, L., Russell, H. R., McKinnon, P. J., Solnica-Krezel, L. & Oliver, G. Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development. Genes Dev 17, 368-79 (2003).

Lazebnik, M. B., Tussie-Luna, M. I. & Roy, A. L. Determination and Functional Analysis of the Consensus Binding Site for TFII-I Family Member BEN, Implicated in Williams-Beuren Syndrome. J Biol Chem 283, 11078-82 (2008).

Lee, H. H. & Frasch, M. Nuclear integration of positive Dpp signals, antagonistic Wg inputs and mesodermal competence factors during Drosophila visceral mesoderm induction. Development 132, 1429-42 (2005).

Lee, T. I. & Young, R. A. Transcription of eukaryotic protein-coding genes. Annu Rev Genet 34, 77-137 (2000).

Levine, H. A. & Nilsen-Hamilton, M. A mathematical analysis of SELEX. Comput Biol Chem 31, 11-35 (2007).

Levine, M. & Davidson, E. H. Gene regulatory networks for development. Proc Natl Acad Sci U S A 102, 4936-42 (2005).

Lewis, E. B. A gene complex controlling segmentation in Drosophila. Nature 276, 565-70 (1978).

Li, L., Liang, Y. & Bass, R. L. GAPWM: a genetic algorithm method for optimizing a position weight matrix. Bioinformatics 23, 1188-94 (2007).

Li, N. & Tompa, M. Analysis of computational approaches for motif discovery. Algorithms Mol Biol 1, 8 (2006).

Li, X., Oghi, K. A., Zhang, J., Krones, A., Bush, K. T., Glass, C. K., Nigam, S. K., Aggarwal, A. K., Maas, R., Rose, D. W. & Rosenfeld, M. G. Eya protein phosphatase activity regulates Six1-Dach-Eya transcriptional effects in mammalian organogenesis. Nature 426, 247-54 (2003).

Lifanov, A. P., Makeev, V. J., Nazina, A. G. & Papatsenko, D. A. Homotypic regulatory clusters in Drosophila. Genome Res 13, 579-88 (2003).

Linnell, J., Mott, R., Field, S., Kwiatkowski, D. P., Ragoussis, J. & Udalova, I. A. Quantitative high-throughput analysis of transcription factor binding specificities. Nucleic Acids Res 32, e44 (2004).

Liu, L. A. & Bader, J. S. Ab initio prediction of transcription factor binding sites. Pac Symp Biocomput, 484-95 (2007).

Loosli, F., Winkler, S. & Wittbrodt, J. Six3 overexpression initiates the formation of ectopic retina. Genes Dev 13, 649-54 (1999).

Lopez-Rios, J., Tessmar, K., Loosli, F., Wittbrodt, J. & Bovolenta, P. Six3 and Six6 activity is modulated by members of the groucho family. Development 130, 185-95 (2003).

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'Hoy, K. & et al. Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. Science 255, 1253-5 (1992).

Mankodi, A., Logigian, E., Callahan, L., McClain, C., White, R., Henderson, D., Krym, M. & Thornton, C. A. Myotonic dystrophy in transgenic mice expressing an expanded CUG repeat. Science 289, 1769-73 (2000).

Marinescu, V. D., Kohane, I. S. & Riva, A. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinformatics 6, 79 (2005).

Marshall, K. A. & Ellington, A. D. In vitro selection of RNA aptamers. Methods Enzymol 318, 193-214 (2000).

Matthews, B. W. Protein-DNA interaction. No code for recognition. Nature 335, 294-5 (1988).

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. & Wingender, E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31, 374-8 (2003).

McGinnis, W., Levine, M. S., Hafen, E., Kuroiwa, A. & Gehring, W. J. A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes. Nature 308, 428-33 (1984).

McGrail, M., Gepner, J., Silvanovich, A., Ludmann, S., Serr, M. & Hays, T. S. Regulation of cytoplasmic dynein function in vivo by the Drosophila Glued complex. J Cell Biol 131, 411-25 (1995).

Metz, C. E. Basic principles of ROC analysis. Semin Nucl Med 8, 283-98 (1978).

Miller, J. M., Oligino, T., Pazdera, M., Lopez, A. J. & Hoshizaki, D. K. Identification of fat-cell enhancer regions in Drosophila melanogaster. Insect Mol Biol 11, 67-77 (2002).

Mitsialis, S. A. & Kafatos, F. C. Regulatory elements controlling chorion gene expression are conserved between flies and moths. Nature 317, 453-6 (1985).

Mohler, J., Eldon, E. D. & Pirrotta, V. A novel spatial transcription pattern associated with the segmentation gene, giant, of Drosophila. Embo J 8, 1539-48 (1989).

Montgomery, S. B., Griffith, O. L., Sleumer, M. C., Bergman, C. M., Bilenky, M., Pleasance, E. D., Prychyna, Y., Zhang, X. & Jones, S. J. ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. Bioinformatics 22, 637-40 (2006).

Moore, L. A., Broihier, H. T., Van Doren, M. & Lehmann, R. Gonadal mesoderm and fat body initially follow a common developmental path in Drosophila. Development 125, 837-44 (1998).

Moore, L. A., Broihier, H. T., Van Doren, M., Lunsford, L. B. & Lehmann, R. Identification of genes controlling germ cell migration and embryonic gonad formation in Drosophila. Development 125, 667-78 (1998).

Mount, D. W. Bioinformatics: Sequence and Genome Analysis (CSHL Press, 2004).

Murre C, M. P., Vaessin H, Caudy M, Jan LY, Jan YN, Cabrera CV, Buskin JN, Hauschka SD, Lassar AB, et al. . Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. cell 58, 537-44 (1989).

Musheev, M. U. & Krylov, S. N. Selection of aptamers by systematic evolution of ligands by exponential enrichment: addressing the polymerase chain reaction issue. Anal Chim Acta 564, 91-6 (2006).

Odenwald, W. F., Rasband, W., Kuzin, A. & Brody, T. EVOPRINTER, a multigenomic comparative tool for rapid identification of functionally important DNA. Proc Natl Acad Sci U S A 102, 14700-5 (2005).

Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., Fraenkel, E., Bell, G. I. & Young, R. A. Control of pancreas and liver gene expression by HNF transcription factors. Science 303, 1378-81 (2004).

Ohto, H., Kamada, S., Tago, K., Tominaga, S. I., Ozaki, H., Sato, S. & Kawakami, K. Cooperation of six and eya in activation of their target genes through nuclear translocation of Eya. Mol Cell Biol 19, 6815-24 (1999).

Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol Cell Biol 9, 2944-9 (1989).

Oliver, G., Mailhos, A., Wehr, R., Copeland, N. G., Jenkins, N. A. & Gruss, P. Six3, a murine homologue of the sine oculis gene, demarcates the most anterior

border of the developing neural plate and is expressed during eye development. Development 121, 4045-55 (1995).

Orengo, D. J. & Aguade, M. Genome scans of variation and adaptive change: extended analysis of a candidate locus close to the phantom gene region in Drosophila melanogaster. Mol Biol Evol 24, 1122-9 (2007).

Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem Sci 25, 99-104 (2000).

Orlando, V., Strutt, H. & Paro, R. Analysis of chromatin structure by in vivo formaldehyde cross-linking. Methods 11, 205-14 (1997).

Ostrin, E. J., Li, Y., Hoffman, K., Liu, J., Wang, K., Zhang, L., Mardon, G. & Chen, R. Genome-wide identification of direct targets of the Drosophila retinal determination protein Eyeless. Genome Res 16, 466-76 (2006).

Ouali, M. & King, R. D. Cascaded multiple classifiers for secondary structure prediction. Protein Sci 9, 1162-76 (2000).

Ozaki, H., Watanabe, Y., Takahashi, K., Kitamura, K., Tanaka, A., Urase, K., Momoi, T., Sudo, K., Sakagami, J., Asano, M., Iwakura, Y. & Kawakami, K. Six4, a putative myogenin gene regulator, is not essential for mouse embryonal development. Mol Cell Biol 21, 3343-50 (2001).

Papatsenko, D. A., Makeev, V. J., Lifanov, A. P., Regnier, M., Nazina, A. G. & Desplan, C. Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers. Genome Res 12, 470-81 (2002).

Paterson, B. M., Shirakata, M., Nakamura, S., Dechesne, C., Walldorf, U., Eldridge, J., Dubendorfer, A., Frasch, M. & Gehring, W. J. Isolation and functional comparison of Dmyd, the Drosophila homologue of the vertebrate myogenic determination genes, with CMD1. Symp Soc Exp Biol 46, 89-109 (1992).

Paterson, B. M., Walldorf, U., Eldridge, J., Dubendorfer, A., Frasch, M. & Gehring, W. J. The Drosophila homologue of vertebrate myogenic-determination genes encodes a transiently expressed nuclear protein marking primary myogenic cells. Proc Natl Acad Sci U S A 88, 3782-6 (1991).

Pettifer, S. R., Sinnott, J. R. & Attwood, T. K. UTOPIA-User-Friendly Tools for Operating Informatics Applications. Comp Funct Genomics 5, 56-60 (2004).

Pham, Y. C., Man, N., Holt, I., Sewry, C. A., Pall, G., Johnson, K. & Morris, G. E. Characterisation of the transcription factor, SIX5, using a new panel of monoclonal antibodies. J Cell Biochem 95, 990-1001 (2005).

Piano, F., Parisi, M. J., Karess, R. & Kambysellis, M. P. Evidence for redundancy but not trans factor-cis element coevolution in the regulation of Drosophila Yp genes. Genetics 152, 605-16 (1999).

Pirrotta, V., Bickel, S. & Mariani, C. Developmental expression of the Drosophila zeste gene and localization of zeste protein on polytene chromosomes. Genes Dev 2, 1839-50 (1988).

Pollock, R. & Treisman, R. A sensitive method for the determination of protein-DNA binding specificities. Nucleic Acids Res 18, 6197-204 (1990).

Ponzielli, R., Astier, M., Chartier, A., Gallet, A., Therond, P. & Semeriva, M. Heart tube patterning in Drosophila requires integration of axial and segmental information provided by the Bithorax Complex genes and hedgehog signaling. Development 129, 4509-21 (2002).

Popichenko, D., Sellin, J., Bartkuhn, M. & Paululat, A. Hand is a direct target of the forkhead transcription factor Biniou during Drosophila visceral mesoderm differentiation. BMC Dev Biol 7, 49 (2007).

Posch, S., Grau, J., Gohr, A., Ben-Gal, I., Kel, A. E. & Grosse, I. Recognition of cis-regulatory elements with vombat. J Bioinform Comput Biol 5, 561-77 (2007).

Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E. D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics 3, 30 (2002).

Rakocevic-Stojanovic, V., Milovanovic, B., Ivic, N., Ille, T., Marjanovic, I., Stevic, Z., Pavlovic, S. & Lavrnic, D. Cardiac autonomic nervous system in patients with myotonic dystrophy type 1. Acta Myol 26, 112-4 (2007).

Rayapureddi, J. P., Kattamuri, C., Steinmetz, B. D., Frankfort, B. J., Ostrin, E. J., Mardon, G. & Hegde, R. S. Eyes absent represents a class of protein tyrosine phosphatases. Nature 426, 295-8 (2003).

Rebay, I., Silver, S. J. & Tootle, T. L. New vision from Eyes absent: transcription factors as enzymes. Trends Genet 21, 163-71 (2005).

Reddy, S., Smith, D. B., Rich, M. M., Leferovich, J. M., Reilly, P., Davis, B. M., Tran, K., Rayburn, H., Bronson, R., Cros, D., Balice-Gordon, R. J. & Housman, D. Mice lacking the myotonic dystrophy protein kinase develop a late onset progressive myopathy. Nat Genet 13, 325-35 (1996).

Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314, 1041-52 (2001).

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. & Young, R. A. Genome-wide location and function of DNA binding proteins. Science 290, 2306-9 (2000).

Rimmele, M. Nucleic acid aptamers as tools and drugs: recent developments. Chembiochem 4, 963-71 (2003).

Ristow, M. Neurodegenerative disorders associated with diabetes mellitus. J Mol Med 82, 510-29 (2004).

Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16, 939-45 (1998).

Ruiz-Gomez, M., Coutts, N., Suster, M. L., Landgraf, M. & Bate, M. myoblasts incompetent encodes a zinc finger transcription factor required to specify fusion-competent myoblasts in Drosophila. Development 129, 133-41 (2002).

Ryan, A. K. & Rosenfeld, M. G. POU domain family values: flexibility, partnerships, and developmental codes. Genes Dev 11, 1207-25 (1997).

Sagel, J., Distiller, L. A., Morley, J. E., Isaacs, H., Kay, G. & Van Der Walt, A. Myotonia dystrophica: Studies on gonadal function using luteinizing hormone-releasing hormone (LRH). J Clin Endocrinol Metab 40, 1110-3 (1975).

Sahly, I., Andermann, P. & Petit, C. The zebrafish eya1 gene and its expression pattern during embryogenesis. Dev Genes Evol 209, 399-410 (1999).

Sandve, G. K. & Drablos, F. A survey of motif discovery methods in an integrated framework. Biol Direct 1, 11 (2006).

Sanggaard, K. M., Rendtorff, N. D., Kjaer, K. W., Eiberg, H., Johnsen, T., Gimsing, S., Dyrmose, J., Nielsen, K. O., Lage, K. & Tranebjaerg, L. Branchio-oto-renal syndrome: detection of EYA1 and SIX1 mutations in five out of six Danish families by combining linkage, MLPA and sequencing analyses. Eur J Hum Genet 15, 1121-31 (2007).

Santos, A. C. & Lehmann, R. Isoprenoids control germ cell migration downstream of HMGCoA reductase. Dev Cell 6, 283-93 (2004).

Sarkar, P. S., Appukuttan, B., Han, J., Ito, Y., Ai, C., Tsai, W., Chai, Y., Stout, J. T. & Reddy, S. Heterozygous loss of Six5 in mice is sufficient to cause ocular cataracts. Nat Genet 25, 110-4 (2000).

Sarkar, P. S., Paul, S., Han, J. & Reddy, S. Six5 is required for spermatogenic cell survival and spermiogenesis. Hum Mol Genet 13, 1421-31 (2004).

Sato, S., Nakamura, M., Cho, D. H., Tapscott, S. J., Ozaki, H. & Kawakami, K. Identification of transcriptional targets for Six5: implication for the pathogenesis of myotonic dystrophy type 1. Hum Mol Genet 11, 1045-58 (2002).

Schlosser, G. & Ahrens, K. Molecular anatomy of placode development in Xenopus laevis. Dev Biol 271, 439-66 (2004).

Schmid, C. D., Perier, R., Praz, V. & Bucher, P. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. Nucleic Acids Res 34, D82-5 (2006).

Scholz, H., Deatrick, J., Klaes, A. & Klambt, C. Genetic dissection of pointed, a Drosophila gene encoding two ETS-related proteins. Genetics 135, 455-68 (1993).

Schug, J. & Overton, G. C. Modeling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding. Proc Int Conf Intell Syst Mol Biol 5, 268-71 (1997).

Scopes, R. K. Strategies for protein purification. Curr Protoc Protein Sci Chapter 1, Unit 1 2 (2001).

Scott, M. P. & Weiner, A. J. Structural relationships among genes that control development: sequence homology between the Antennapedia, Ultrabithorax, and fushi tarazu loci of Drosophila. Proc Natl Acad Sci U S A 81, 4115-9 (1984).

Seeman, N. C., Rosenberg, J. M. & Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A 73, 804-8 (1976).

Seimiya, M. & Gehring, W. J. The Drosophila homeobox gene optix is capable of inducing ectopic eyes by an eyeless-independent mechanism. Development 127, 1879-86 (2000).

Seo, H. C., Curtiss, J., Mlodzik, M. & Fjose, A. Six class homeobox genes in drosophila belong to three distinct families and are involved in head development. Mech Dev 83, 127-39 (1999).

Shamah, S. M., Healy, J. M. & Cload, S. T. Complex target SELEX. Acc Chem Res 41, 130-8 (2008).

Sharan, R., Ben-Hur, A., Loots, G. G. & Ovcharenko, I. CREME: Cis-Regulatory Module Explorer for the human genome. Nucleic Acids Res 32, W253-6 (2004).

Sharan, R., Ovcharenko, I., Ben-Hur, A. & Karp, R. M. CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. Bioinformatics 19 Suppl 1, i283-91 (2003).

Shaw, D. J., McCurrach, M., Rundle, S. A., Harley, H. G., Crow, S. R., Sohn, R., Thirion, J. P., Hamshere, M. G., Buckler, A. J., Harper, P. S. & et al. Genomic organization and transcriptional units at the myotonic dystrophy locus. Genomics 18, 673-9 (1993).

Shi, Y. & Shi, Y. Metabolic enzymes and coenzymes in transcription--a direct link between metabolism and transcription? Trends Genet 20, 445-52 (2004).

Shimell, M. J., Peterson, A. J., Burr, J., Simon, J. A. & O'Connor, M. B. Functional analysis of repressor binding sites in the iab-2 regulatory region of the abdominal-A homeotic gene. Dev Biol 218, 38-52 (2000).

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W. & Haussler, D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15, 1034-50 (2005).

Silver, S. J. & Rebay, I. Signaling circuitries in development: insights from the retinal determination gene network. Development 132, 3-13 (2005).

Smith, D. B. & Johnson, K. S. Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase. Gene 67, 31-40 (1988).

Snyder, M., Huang, X. Y. & Zhang, J. J. Identification of novel direct Stat3 target genes for control of growth and differentiation. J Biol Chem 283, 3791-8 (2008).

Spitz, F., Demignon, J., Porteu, A., Kahn, A., Concordet, J. P., Daegelen, D. & Maire, P. Expression of myogenin during embryogenesis is controlled by Six/sine oculis homeoproteins through a conserved MEF3 binding site. Proc Natl Acad Sci U S A 95, 14220-5 (1998).

Staehling-Hampton, K. & Hoffmann, F. M. Ectopic decapentaplegic in the Drosophila midgut alters the expression of five homeotic genes, dpp, and wingless, causing specific morphological defects. Dev Biol 164, 502-12 (1994).

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S. W., Han, M. V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J.,

Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M. & Kellis, M. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450, 219-32 (2007).

Starz-Gaiano, M. & Lehmann, R. Moving towards the next generation. Mech Dev 105, 5-18 (2001).

Strutt, H., Cavalli, G. & Paro, R. Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. Embo J 16, 3621-32 (1997).

Suzuki-Yagawa, Y., Kawakami, K. & Nagano, K. Housekeeping Na,K-ATPase alpha 1 subunit gene promoter is composed of multiple cis elements to which common and cell type-specific factors bind. Mol Cell Biol 12, 4046-55 (1992).

Tanaka, K. Myotonic dystrophy. Med Hypotheses 17, 415-25 (1985).

Tanaka, K., Takeshita, K. & Takita, M. Deoxycholic acid, a candidate for the maternal intrauterine factor in early-onset myotonic dystrophy. Lancet 1, 1046-7 (1981).

Tanaka, Y., Suzuki, Y., Shimozawa, N., Nanba, E. & Kondo, N. Congenital myotonic dystrophy: report of paternal transmission. Brain Dev 22, 132-4 (2000).

Tapanes-Castillo, A. & Baylies, M. K. Notch signaling patterns Drosophila mesodermal segments by regulating the bHLH transcription factor twist. Development 131, 2359-72 (2004).

Tapscott, S. J. Deconstructing myotonic dystrophy. Science 289, 1701-2 (2000).

Tavsanli, B. C., Ostrin, E. J., Burgess, H. K., Middlebrooks, B. W., Pham, T. A. & Mardon, G. Structure-function analysis of the Drosophila retinal determination protein Dachshund. Dev Biol 272, 231-47 (2004).

Taylor, W. R. Transcription and translation in an RNA world. Philos Trans R Soc Lond B Biol Sci 361, 1751-60 (2006).

Thiesen, H. J. & Bach, C. Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. Nucleic Acids Res 18, 3203-9 (1990).

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P. & Moreau, Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 9, 447-64 (2002).

Thisse, B., el Messal, M. & Perrin-Schmitt, F. The twist gene: isolation of a Drosophila zygotic gene necessary for the establishment of dorsoventral pattern. Nucleic Acids Res 15, 3439-53 (1987).

Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. & van Helden, J. RSAT: regulatory sequence analysis tools. Nucleic Acids Res (2008).

Thornton, C. A., Wymer, J. P., Simmons, Z., McClain, C. & Moxley, R. T., 3rd. Expansion of the myotonic dystrophy CTG repeat reduces expression of the flanking DMAHP gene. Nat Genet 16, 407-9 (1997).

Tokgozoglu, L. S., Ashizawa, T., Pacifico, A., Armstrong, R. M., Epstein, H. F. & Zoghbi, W. A. Cardiac involvement in a large kindred with myotonic dystrophy. Quantitative assessment and relation to size of CTG repeat expansion. Jama 274, 813-9 (1995).

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S. E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S. E. & Rubin, G. M. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3, RESEARCH0088 (2002).

Tomotsune, D., Shoji, H., Wakamatsu, Y., Kondoh, H. & Takahashi, N. A mouse homologue of the Drosophila tumour-suppressor gene l(2)gl controlled by Hox-C8 in vivo. Nature 365, 69-72 (1993).

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. & Zhu, Z. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23, 137-44 (2005).

Tootle, T. L., Silver, S. J., Davies, E. L., Newman, V., Latek, R. R., Mills, I. A., Selengut, J. D., Parlikar, B. E. & Rebay, I. The transcription factor Eyes absent is a protein tyrosine phosphatase. Nature 426, 299-302 (2003).

Torres, F. A. & Bonner, J. J. Genetic identification of the site of DNA contact in the yeast heat shock transcription factor. Mol Cell Biol 15, 5063-70 (1995).

Toy, J. & Sundin, O. H. Expression of the optx2 homeobox gene during mouse development. Mech Dev 83, 183-6 (1999).

Treisman, J., Harris, E. & Desplan, C. The paired box encodes a second DNA-binding domain in the paired homeo domain protein. Genes Dev 5, 594-604 (1991).

Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249, 505-10 (1990).

Turnpenny, P., Clark, C. & Kelly, K. Intelligence quotient profile in myotonic dystrophy, intergenerational deficit, and correlation with CTG amplification. J Med Genet 31, 300-5 (1994).

Van Doren, M. Development of the Somatic Gonad and Fat Bodies (ed. H., S.) (Springer, New York, 2006).

Van Dyke, M. W., Hertzberg, R. P. & Dervan, P. B. Map of distamycin, netropsin, and actinomycin binding sites on heterogeneous DNA: DNA cleavage-inhibition patterns with methidiumpropyl-EDTA.Fe(II). Proc Natl Acad Sci U S A 79, 5470-4 (1982).

Van Dyke, M. W., Van Dyke, N. & Sunavala-Dossabhoy, G. REPSA: general combinatorial approach for identifying preferred ligand-DNA binding sequences. Methods 42, 118-27 (2007).

van Helden, J. Regulatory sequence analysis tools. Nucleic Acids Res 31, 3593-6 (2003).

van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281, 827-42 (1998).

van Helden, J., Rios, A. F. & Collado-Vides, J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 28, 1808-18 (2000).

Vant-Hull, B., Payano-Baez, A., Davis, R. H. & Gold, L. The mathematics of SELEX against complex targets. J Mol Biol 278, 579-97 (1998).

Venkatesh, T. V., Park, M., Ocorr, K., Nemaceck, J., Golden, K., Wemple, M. & Bodmer, R. Cardiac enhancer activity of the homeobox gene tinman depends on CREB consensus binding sites in Drosophila. Genesis 26, 55-66 (2000).

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., van Roy, F. & Lenhard, B. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res 34, D95-7 (2006).

Wakimoto, H., Maguire, C. T., Sherwood, M. C., Vargas, M. M., Sarkar, P. S., Han, J., Reddy, S. & Berul, C. I. Characterization of cardiac conduction system abnormalities in mice with targeted disruption of Six5 gene. J Interv Card Electrophysiol 7, 127-35 (2002).

Wallis, D. E., Roessler, E., Hehr, U., Nanni, L., Wiltshire, T., Richieri-Costa, A., Gillessen-Kaesbach, G., Zackai, E. H., Rommens, J. & Muenke, M. Mutations in

the homeodomain of the human SIX3 gene cause holoprosencephaly. Nat Genet 22, 196-8 (1999).

Wang, Y. H. Chromatin structure of repeating CTG/CAG and CGG/CCG sequences in human disease. Front Biosci 12, 4731-41 (2007).

Wang, Y. H., Amirhaeri, S., Kang, S., Wells, R. D. & Griffith, J. D. Preferential nucleosome assembly at DNA triplet repeats from the myotonic dystrophy gene. Science 265, 669-71 (1994).

Wansink, D. G. & Wieringa, B. Transgenic mouse models for myotonic dystrophy type 1 (DM1). Cytogenet Genome Res 100, 230-42 (2003).

Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5, 276-87 (2004).

Wawersik, S. & Maas, R. L. Vertebrate eye development as modeled in Drosophila. Hum Mol Genet 9, 917-25 (2000).

Wei, W. & Yu, X. D. Comparative analysis of regulatory motif discovery tools for transcription factor binding sites. Genomics Proteomics Bioinformatics 5, 131-42 (2007).

Wellik, D. M., Hawkes, P. J. & Capecchi, M. R. Hox11 paralogous genes are essential for metanephric kidney induction. Genes Dev 16, 1423-32 (2002).

Wheeler, T. M. & Thornton, C. A. Myotonic dystrophy: RNA-mediated muscle disease. Curr Opin Neurol 20, 572-6 (2007).

Wilson, D. S., Guenther, B., Desplan, C. & Kuriyan, J. High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. Cell 82, 709-19 (1995).

Winchester, C. L., Ferrier, R. K., Sermoni, A., Clark, B. J. & Johnson, K. J. Characterization of the expression of DMPK and SIX5 in the human eye and implications for pathogenesis in myotonic dystrophy. Hum Mol Genet 8, 481-92 (1999).

Wolberger, C. Homeodomain interactions. Curr Opin Struct Biol 6, 62-8 (1996).

Wood, W. B. Host specificity of DNA produced by Escherichia coli: bacterial mutations affecting the restriction and modification of DNA. J Mol Biol 16, 118-33 (1966).

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. & Romano, L. A. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20, 1377-419 (2003).

Wright, W. E., Binder, M. & Funk, W. Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. Mol Cell Biol 11, 4104-10 (1991).

Xiang, M., Zhou, L., Macke, J. P., Yoshioka, T., Hendry, S. H., Eddy, R. L., Shows, T. B. & Nathans, J. The Brn-3 family of POU-domain factors: primary structure, binding specificity, and expression in subsets of retinal ganglion cells and somatosensory neurons. J Neurosci 15, 4762-85 (1995).

Xie, D., Cai, J., Chia, N. Y., Ng, H. H. & Zhong, S. Cross-species de novo identification of cis-regulatory modules with GibbsModule: application to gene regulation in embryonic stem cells. Genome Res 18, 1325-35 (2008).

Xu, P. X., Adams, J., Peters, H., Brown, M. C., Heaney, S. & Maas, R. Eya1-deficient mice lack ears and kidneys and show abnormal apoptosis of organ primordia. Nat Genet 23, 113-7 (1999).

Xu, P. X., Woo, I., Her, H., Beier, D. R. & Maas, R. L. Mouse Eya homologues of the Drosophila eyes absent gene require Pax6 for expression in lens and nasal placode. Development 124, 219-31 (1997).

Xu, P. X., Zheng, W., Huang, L., Maire, P., Laclef, C. & Silvius, D. Six1 is required for the early organogenesis of mammalian kidney. Development 130, 3085-94 (2003).

Xue, L. & Noll, M. Dual role of the Pax gene paired in accessory gland development of Drosophila. Development 129, 339-46 (2002).

Yamada, Y., Davis, K. D. & Coffman, C. R. Programmed cell death of primordial germ cells in Drosophila is regulated by p53 and the Outsiders monocarboxylate transporter. Development 135, 207-16 (2008).

Yang, E., Simcha, D., Almon, R. R., Dubois, D. C., Jusko, W. J. & Androulakis, I. P. Context specific transcription factor prediction. Ann Biomed Eng 35, 1053-67 (2007).

Yanowitz, J. L., Shakir, M. A., Hedgecock, E., Hutter, H., Fire, A. Z. & Lundquist, E. A. UNC-39, the C. elegans homolog of the human myotonic dystrophy-associated homeodomain protein Six5, regulates cell motility and differentiation. Dev Biol 272, 389-402 (2004).

Yavatkar, A. S., Lin, Y., Ross, J., Fann, Y., Brody, T. & Odenwald, W. F. Rapid detection and curation of conserved DNA via enhanced-BLAT and EvoPrinterHD analysis. BMC Genomics 9, 106 (2008).

Yin, Z., Xu, X. L. & Frasch, M. Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. Development 124, 4971-82 (1997).

Yoshida, T. MCAT elements and the TEF-1 family of transcription factors in muscle development and disease. Arterioscler Thromb Vasc Biol 28, 8-17 (2008).

Zheng, W., Huang, L., Wei, Z. B., Silvius, D., Tang, B. & Xu, P. X. The role of Six1 in mammalian auditory system development. Development 130, 3989-4000 (2003).

Zhu, C. C., Dyer, M. A., Uchikawa, M., Kondoh, H., Lagutin, O. V. & Oliver, G. Six3-mediated auto repression and eye development requires its interaction with members of the Groucho-related family of co-repressors. Development 129, 2835-49 (2002).

Zou, D., Silvius, D., Fritzsch, B. & Xu, P. X. Eya1 and Six1 are essential for early steps of sensory neurogenesis in mammalian cranial placodes. Development 131, 5561-72 (2004).

Zuber, M. E., Perron, M., Philpott, A., Bang, A. & Harris, W. A. Giant eyes in Xenopus laevis by overexpression of XOptx2. Cell 98, 341-52 (1999).

Zweig, M. H. & Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39, 561-77 (1993).

Table of Contents