CARDIFF
UNIVERSITY
PRIFYSGOL
CAERDYDD

Functional analysis of the promoter regions of *de novo* genes_090 & 074 in *Drosophila melanogaster* and *in vitro* Topi C-terminal-DNA interaction by SELEX

Shrinivas Nivrutti Dighe

December 2019

This thesis is submitted for the

Degree of Doctor of Philosophy

This thesis is dedicated to my

Mum & Dad

Acknowledgements

I would like to thank my supervisor Helen White-Cooper for her guidance, patience and enthusiasm throughout my PhD. It has been a wonderful journey in the last four years and I am grateful for all the help I have received during my projects. I would also like to thank Nick Kent and Angela Marchbank for providing guidance for SELEX-sequencing. I would also like to thank all the PIs in the *Drosophila* facility, Sonia Lopez de Quinto, Mike Taylor, Wynand van der Goes van Naters, Joaquin de Navascués Melero, Gaynor Smith, Owen Peters, and Fisun Hamaratoglu for all their valuable feedback during my PhD.

I would like to express my gratitude to my fellow colleagues in the fly lab. My heartfelt thanks to my friends Swati, Santosh, Prashant, Bhushan and Vina for their support and motivation throughout my PhD.

My deepest thanks to my Mum & Dad for your love and encouragement, also to my siblings Vikas, Punam, Preety and their families for their love and support.

Summary

de novo genes are derived from ancestral non-coding sequences, and typically show strongly testis biased expression. Sequence variation in cis-acting regulatory regions presumably generates sites that are recognised by testis-specific transcriptional regulators, such as tMAC. The causative mutations in *de novo* gene evolution can be identified by comparing naturally expressing and non-expressing alleles of not-yet-fixed de novo genes. Promoters of expressed and non-expressed alleles differ by the presence of SNPs and/or indels.

In the first part of my thesis, I have investigated the molecular mechanisms of testis-specific expressed *de novo* genes_090 & 074. For gene_090, reporters showed that a 7bp deletion just upstream of the TSS is necessary and sufficient to convert a natural non-expressing allele into an expressing allele. Further constructs indicate the deletion in expressed allele does not create any binding site for regulatory proteins per se but the mutation of the two base pairs flanking the deletion site suggests that the sequence created by the deletion is critical, as well as the spacing of other flanking sequences are important for expression. Expression of 074 has evolved independently at least twice. Two expressing alleles do not share any SNPs or indels not also found in at least one non-expressing allele, yet both are able to drive expression of a reporter gene. Intriguingly, for both 090 and 074, the reporter mRNA is translationally repressed, and only translated in late spermatids, reminiscent of many more ancient testis-specific transcripts.

In the second part of my thesis, I have discovered the *de novo* motif of 15bp for Topi C-terminal protein. The discovered motif was characterised by EMSA and *in silico* analysis observed the occurrence of discovered motif in promoter regions of *aly* dependent genes. Altogether, these results indeed support the motif being real.

Table of Contents

Abbreviations

BCIP	X- phosphate/5-Bromo-4-chloro-3-indolyl-phosphate
DIC	Differential Interference Contrast
DIG	Digoxygenin
dREAM	Drosophila Rb, E2F and Myb
DREAM	DP, Rb and MuvB
EDTA	Ethylenediamine tetra-acetic acid
EMSA	Electrophoretic mobility shift assay
HB	Hybridisation buffer
HDACs	Histone deacetylase
His	Histidine
HP	High pH
IPTG	Isopropyl-beta-D-thiogalactopyranoside
LB	Luria-Bertani
Ni-NTA	Nickel-nitrilotriacetic acid
NBT	Nitro blue tetrazolium chloride
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PBS-T	Phosphate buffered saline-Tween
PcG	Polycomb group
PCR	Polymerase chain reaction
SDS	Sodium dodecyl sulphate
TAE	Tris-acetate-Ethylenediaminetetraacetic acid

TFs	Transcription Factors
TCE	Translational control element
tMAC	Testis-specific meiotic arrest complex
TRE	Translational repression element
TSS	Transcription start site
tTAFs	testis TATA binding protein associated factors
UTR	Untranslated region
WT	Wild type
Y2H	Yeast two hybrid

List of Figures

List of Tables

Chapter 1. Introduction

Spermatogenesis is a very specialised and dynamic process by which a male animal developmentally regulates the production of sperm from diploid germline stem cells. Spermatogenesis is the paradigmatic system to study gene expression regulation because of the disciplined transcriptional program in the testis of the particular organism. For our studies we have selected *Drosophila* spermatogenesis as a model system and its genetics allow us to do the various manipulations in the *Drosophila* genome.

1.1. *Drosophila* Spermatogenesis

Spermatogenesis occurs in the male's gonads and it is the process of sperm production which involves meiotic as well as mitotic cellular division. It is an essential process in almost all of the organisms and it needs to be tightly regulated throughout the process for mature sperm production. Spermatogenesis contributes to the maintenance of the species by producing a sperm which further fuses with the ovum and gives rise to a whole new organism through intense cellular divisions and extensive morphological changes.

1.1.1. Germline proliferation centre

In *Drosophila*, a testis is a structurally blunt ended, coiled tube, and place where sperm production occurs throughout the life of a fly (Demarco et al. 2014). At the apical tip of the testis is the hub, a group of post-mitotic somatic cells which are responsible for the maintenance of the germline and stem cells through cellular signaling (Sinden et al. 2012, Singh et al. 2010). The hub secretes the ligand Unpaired (Upd) which activates a JAK-STAT signaling pathway in the adjacent stem cells **(Figure 1.1)** (Sinden et al. 2012, Kiger et al. 2001). On average each mitotic division of germline cells (GSC) produces one GSC and one spermatogonium (Tulina and Matunis 2001). Similarly, each cyst stem cell (CySC) (formerly known as cyst progenitor cell) division gives rise to CySC and cyst cell (Voog, D'Alterio, and Jones 2008). Two post mitotic cyst cells surround each spermatogonium and nourish it, analogous to the Sertoli cells in mammals (White-Cooper 2010).

1.1.2. Mitosis

Both GSC and CySC divide asymmetrically and produce gonialblasts and cyst cells respectively that form developmental unit, called a cyst. In each cyst four rounds of synchronous mitotic division happen developing in 2, 4, 8, and 16 cell syncytia of spermatogonial cells. Mitotic divisions are kept limited to four by secretion of two signaling molecules (Gbb and Dpp are required for maintenance of male GSCs) by cyst cell and hub cells and activate a Bone morphogenetic protein (BMP) family signalling by regulating the expression of two genes *bag of marbles (bam)* and *benign gonial cell neoplasm (bgcn)* (Kawase et al. 2004). Further these two genes *bam* and *bgcn* control the proliferation of amplifying germ cells and play a central role in switching the program from mitosis to meiosis (Ohlstein et al. 2000, Zhao et al. 2013). Cytokinesis in spermatogonial mitosis and during subsequent meiotic divisions, ceases incompletely, and the contractile rings convert into cytoplasmic bridges called canal bridges. The sharing of cytoplasm through these ring canals plays a vital role in synchronizing the cell cycle and differentiation of cyst cells (Spradling et al. 2011).

1.1.3. Spermatocytes

Premeiotic S-phase occurs early in the primary spermatocytes and then developing primary spermatocytes begin an extended G2 phase consisting of rapid growth and transcription which lasts approximately 3.5 days during which the volume of cells increase by approximately 25 times and because transcription is primarily shut off after this point (Olivieri and Olivieri 1965) huge number of genes must be transcribed so their mRNAs are present for later stages of spermatogenesis. RNAs needed to generate proteins which are not required until sperm individualisation need to be transcribed and stored. This process should be tightly regulated to guarantee that everything is prepared before the cell enters the meiotic divisions. This coordinated regulation is controlled in-part by the meiotic arrest genes (see the section 1.4.1).

1.1.4. Meiosis

Meiosis involves two linked cell divisions with no intervening S phase. In meiosis-I the homologous chromosome are segregated and then sister chromatids are segregated in meiosis-II resulting in cysts of 64 haploid round spermatids (White-Cooper 2010). The successive progression of the meiotic cell cycle is partly controlled by some of the core cell cycle machinery. For example, the G2/M transition in mitosis and this phase in male meiosis-I is regulated by *CDC2 kinase* (Sigrist et al. 1995) and *Cyclin B* is also needed for the G2/M transition in both mitosis and meiosis. The CDC2/Cyclin B complex needed for G2/M transition is activated by cdc25 phosphatase homologue *string* in mitosis but in a meiosis specific homologue is required, *twine* (Alphey et al. 1992, Courtot et al. 1992, White-Cooper et al. 1998, White-Cooper, Alphey, and Glover 1993). The meiotic G2/M transition also requires *boule* which is a homologue of human *Deleted in Azoospermia (DAZ,* r in sperm production) (Eberhart, Maines, and Wasserman 1996). Activity of boule is required for efficient translation of *twine* mRNA and this was tested by making deficient for *boule* and *twine.* Expression of *twine* reduced dramatically in *boule* mutant background and this reduction occurred at protein level, not at mRNA or accumulation (Maines and Wasserman 1999).

Drosophila male meiosis is different from female meiosis due to lack of the formation of a synaptonemal complex, thus there is no meiotic recombination in male germline (Hawley 2002)**.** The synaptonemal complex (SC) is a highly conserved meiotic structure and five SC proteins have been identified: C(3)G, C(2)M, CONA, ORD, and Corolla. The SC is required for meiotic recombination in *Drosophila* and absence of these protein leads to reduced crossing over and chromosomal nondisjunction (Hemmer and Blumenstiel 2016).

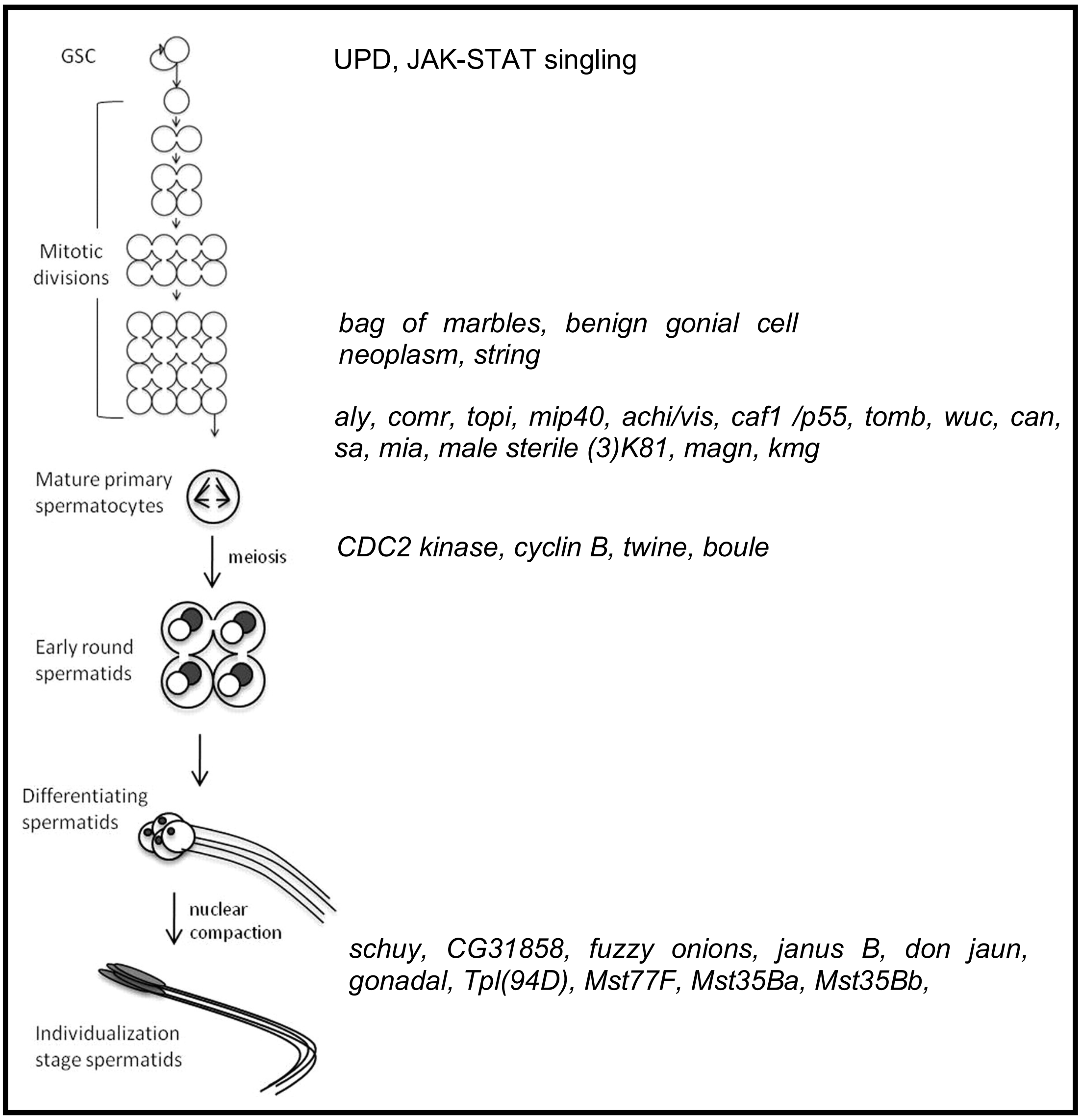

GSC
UPD, JAK-STAT singling
Mitotic divisions
bag of marbles, benign gonial cell neoplasm, string
aly, comr, topi, mip40, achi/vis, caf1 /p55, tomb, wuc, can, sa, mia, male sterile (3)K81, magn, kmg
Mature primary spermatocytes
meiosis
CDC2 kinase, cyclin B, twine, boule
Early round spermatids
Differentiating spermatids
nuclear compaction
schuy, CG31858, fuzzy onions, janus B, don jaun, gonadal, Tpl(94D), Mst77F, Mst35Ba, Mst35Bb,
Individualization stage spermatids

Figure 1.1:- *Drosophila* spermatogenesis and genes required/expressed at various stages for progression of spermatogenesis (Adapted from Sartain *et al.,* 2011).

Immediately after meiosis the mitochondria aggregate and fuse forming a phase-dark mass on the side of the nucleus called the Nebenkern. This stage is generally known as the “onion stage” because in cross section the concentric rings of mitochondrial membrane look like an onion slice.

1.1.5. Spermiogenesis

Spermiogenesis is a process of maturation of spermatids to sperms. Spermiogenesis can progress independently of meiotic cell cycle progression. Specifically, in mutants in which the meiotic divisions are blocked (e.g., *twine*) spermatid differentiation occurs relatively normally (except of course the cells are 4N rather than 1N and have 4 basal bodies so 4 axonemes, and males show sterile phenotype) (Alphey et al. 1992). Round spermatids in the cyst are interconnected, so that they differentiate in synchrony. The major morphological changes of spermatid differentiation include the Nebenkern dissociating into one major and one minor mitochondrial derivative. During the elongation of each spermatid, the nuclei cluster at the head ends of the spermatids, and are oriented towards the basal region of the testis. Flagellar elongation occurs at the tail end of the spermatid and these growing tails push up the testis towards the apical region (Fabian and Brill 2012).

1.1.5.1. Nuclear reshaping and chromatin condensation

During the late stages of elongation, the spherical nuclear mass undertakes a dramatical morphological transformation to become a thin needle like structure only 9µm in length and 0.3µm in width. This process is partly mediated by microtubule dynamics which help remodel the nucleus through a canoe-shape intermediate (Demarco et al. 2014, Kanippayoor, Alpern, and Moehring 2013, Phillips 1970, Farkas et al. 2003, Awe and Renkawitz-Pohl 2010). In mature sperm, chromatin is highly condensed and transcriptionally silent. In *Drosophila,* the sperm nuclear volume decreases to 200 times less than that of somatic cell nucleus. Like mammals, *Drosophila* have two protamine-like genes (*Mst35Ba* and *Mst35Bb*) which represent the chromosomal proteins in mature spermatozoa (Jayaramaiah Raja and Renkawitz-Pohl 2005) and lately a single *Drosophila* transition protein, Tpl(94D) has been identified (Rathke et al. 2007). In nucleus, the chromatin reorganises and histones get replaced by transition protein (Tpl) and then by protamines and Mst77F (Fabian and Brill 2012). During this transition from histone to protamine, histone undergo various modifications (acetylation, ubiquitination), other proteins become sumoylated, transient DNA breaks occur in DNA strands and proteasome activity increases. Histone modifications facilitate access of chromatin remodelling proteins and enzymes. After histone removal, Tpl(94D) is incorporated and together with DNA breaks, may facilitate chromatin unwinding and assembly of protamine and Mst77 on the condensing DNA (Fabian and Brill 2012).

1.1.5.2. Sperm individualisation

The final step of spermiogenesis is the individualisation of the 64 spermatozoa within a cyst. The individualisation complex (IC) is formed with 64 actin-rich cones and migrate from the sperm head along the length of the spermatids (Noguchi and Miller 2003, Fabrizio et al. 1998). At this stage, most of the cytoplasm gets extruded from the cell and cytoplasmic connection between the germline cells within a cyst are lost and remodelling of membrane occurs with all excess material filling a waste bag. This process is facilitated by apoptosis-like process involving elevated caspase activity (Arama, Agapite, and Steller 2003). After this step the mature sperm coils and move eventually to the seminal vesicle, becoming individual motile sperm stored until copulation (White-Cooper 2010).

1.2. The significance of transcription regulation

One of the most intriguing and studied questions in biology is the one related to transference of biological information starting from DNA in multicellular organisms: how is it possible that a single cell with a genome has the potential to develop different tissues, organs and systems? Now we know how this information flows through different biochemical processes, starting from DNA to DNA (replication), from DNA to RNA (transcription) and from RNA to proteins (translation). The first evidence of gene regulation was brought to light by Jacob and Monad and they demonstrated that the protein synthesis, starting from RNA transcription, was mediated by a special type of biomolecules that they further called repressors and these repressors could regulate the gene expression by binding to the specific short DNA sequences (called operators) located in close proximity of the genes (Jacob and Monod 1961). This finding opened up the new research field in molecular biology called transcriptional regulation and shed light on the understanding of gene expression regulation showing (at that time) that gene expression was facilitated by a group of proteins (repressors) and short DNA sequences (operators).

Later, it was found that there was another class of regulatory protein, the activator, in contrast to repressors, they can positively regulate the gene expression (Busby and Ebright 1999, Ma 2011, Zubay, Schwartz, and Beckwith 1970) Now both are commonly known as Transcription factors (TFs). Earlier, most experiments were performed in bacteria, the scientist discovered that the DNA sequences where TFs bind, were located upstream of the gene, a region known as a promoter.

However, although the information flux looks very simple, we should consider that animal genomes develop complex cellular states that give rise to different tissues and in *Drosophila* it gives rise to tissues like, brain, muscles, gut, fat body, antenna etc, with specialised function and this makes us set another question: how cell differentiation is coordinated to give rise to different tissues. Chintapalli *et al*. showed that 67% of all the protein coding genes in the genome are expressed in each tissue (Chintapalli, Wang, and Dow 2007).

A simple answer could be, the gene repertoire; but once again this makes us set more questions like how these genes are regulated?: are all the genes active at the same time? Now we know that the differential gene expression during the development is what gives rise to different types of tissues, cell lines, organs, and homeostasis and the dysregulation of gene expression could be associated with the diseases (Mathelier, Shi, and Wasserman 2015). For example, in cancers, loss of gene expression occurs by transcription silencing introduced by promoter hyper-methylation of CpG islands. In colorectal cancer, colon tumours compared with neighbouring normal-appearing colonic mucosa showed about 600 to 800 heavily methylated CpG islands in promoters of genes in the tumour whereas these CpG islands were not methylated in neighbouring mucosa (Illingworth et al. 2010, Wei et al. 2016).

The answers to these questions could be: a combination of several factors including proteins, RNAs or the DNA itself are responsible for the gene expression regulation. Each step of gene regulation is regulated by various factors and even regulated at different compartments of the cell. One of these steps, called epigenetics, involve chemical modification on DNA or proteins associated with DNA (e.g. histones) that does not alter the sequence of DNA but alters the gene expression or DNA accessibility. Some famous examples are DNA methylation and acetylation, sumoylation, methylation, and phosphorylation to histone-tails, and this then modifies the 3D structure of the chromatin (Wagner et al. 2014, Rossetto, Avvakumov, and Côté 2012, Zhu and Wani 2010, Biterge and Schneider 2014, Sabari et al. 2017, Zhao, Zhang, and Li 2018).

In this thesis I will focus on the transcriptional regulation especially by TFs and the cis and trans regulatory elements, but the reader must not forget that there are other layers of gene expression regulation like mRNA translation and post-translational modification.

To synopsise, cells can sense the external as well as internal stimuli which could affect the gene expression to adapt to various changes. At molecular level, gene expression is regulated by different elements, two of them, the TFs and cis-regulatory elements drive the expression of a particular gene at transcriptional level. The changes in expression goes from activation or inactivation of a gene, to complex processes as cell differentiation (e.g. spermatogenesis) or organogenesis.

1.3. An overview of DNA transcription

Transcription is well-defined as the process which involves the RNA synthesis from a DNA template, the product, mRNA will be later used as a template by the ribosome to synthesise the protein. In any living organism (from bacteria to metazoan), the transcription has been carried out by a special protein complex called RNA-polymerase (RNAP) which is recruited to short sequences located in the promoter. In eukaryotes specific motifs, such as TATA box, work in unison with other elements to recruit RNA polymerase. The promoters are capable of initiating the transcription of associated genes in all cells, using ubiquitously expressed activators are called constitutive promoters. Regulated promoters need the help of TFs or other proteins to recruit or stabilize subunits of the RNAP and turn on or off the expression of a gene (Qin et al. 2010).

Although the first ever studied cis-regulatory elements (CREs) were the promoters, other classes of regulatory elements discovered were called enhancers. These regions can distantly activate gene expression (Banerji, Rusconi, and Schaffner 1981) by contrast to the promoters that would do it locally. Enhancers can be found upstream, downstream, within the introns, or even relatively far away from the genes they regulate.

This discovery also revealed other cis-regulatory elements acting distantly in deactivating the gene expression (silencers) or other sequences capable of avoiding the enhancer activity (insulator). Silencers are CREs that can bind transcription regulation factors (proteins) called repressors, thereby preventing transcription of a gene.

However, we shall not forget that the interactions of these elements are dependent on the accessibility for the regulatory proteins to these cis-regulatory elements, see table 1 for a summary of the regulatory elements responsible for the transcriptional initiation (Lenhard, Sandelin, and Carninci 2012).

Elements	Description	Class
RNA-polymerase (RNAP)	Complex of proteins that transcribe the DNA to RNA	Protein
Transcription Factor (TF)	DNA-binding proteins regulators of gene expression	Protein
Transcription Start Sites (TSS)	First nucleotide transcribed by the RNAP	DNA
Cis-regulatory Module (CRM)	A DNA region with a high concentration of binding sites for distinct TFs	DNA
Transcription Factor Binding Site (TFBS)	A short DNA sequences bound by a TFs	DNA
Enhancer	Distal regulatory (activation) region	DNA
Silencer	Distal regulatory (repression) region	DNA
Insulator	Boundary between distinct chromatin regions. prevent CRMs from one side of the insulator from interacting with TSSs on the other side.	DNA

Table 1:- Classification of transcriptional regulatory elements

1.3.1. Generalities of chromatin structure and histone modifications

The genome needs to be efficiently packed into a small volume in order to fit into the nucleus of a cell. In eukaryotic cells, the large-scale 3D organization of the genome consists of the so-called chromosome territories, the specific region of the nucleus occupied by a chromosome. At intermediate scale, the DNA is folded and form Topological-associated domains (0.5-1Mb) and within them, the DNA can form smaller loops (hundreds of kilobases) (Bonev and Cavalli 2016, Rao et al. 2014, Stevens et al. 2017). A topological associated domain (TAD) is a self-reacting genomic region, meaning that DNA sequences within a TAD physically interacts with each other more frequently that with sequences outside the TAD. Mammalian TAD boundaries are enriched in CCCTC-binding factor (CTCF) sites and TADs are formed by dynamic cohesion-driven loop extrusion and CTCF sites (Williamson et al. 2019). Some studies showed that deletion or inversion of CTCF sits at TAD boundaries can promote TAD boundary talk and rewire enhancer-promoter contacts (Guo et al. 2015, Narendra et al. 2015). To study the CTCF mediated function of TADs in developmental gene regulation, Williamson *et al*., studied the sonic hedgehog (*Shh*) regulatory domain- a locus for long range regulation (Williamson et al. 2019). The SHH morphogen controls the growth and patterning of brain, neural tube and limbs. *Shh* expression regulated by tissue-specific enhancers located within the gene, and upstream in a large gene desert and within neighbouring genes. *Shh* and its *cis*-regulatory elements are all contained within a well characterised ~960kb TAD and loss of CTCF sites at the *Shh* TAD boundaries disrupts chromatin architecture (Williamson et al. 2016). Insulator binding proteins (IBPs) are key players in ensuring the specificity of gene regulation in flies and mammals (Gambetta and Furlong 2018). Of known IBPs, only CTCF site is conserved in both flies and mammals. The CTCF is thought to exert this insulator activity by creating chromatin loops between bound CTCF sites, which prevents physical and regulatory contacts between chromosomal regions that are within loop with those that are outside (Narendra et al. 2015). The presence and orientation of CTCF sites is important for the functionality of these elements (Guo et al. 2015). The CTCF binds to many sites in *Drosophila* genome and is important in regulation of *Hox* genes (Gerasimova et al. 2007). Gambetta *et al*., showed that CTCF is essential for the viability of adult *Drosophila* but importantly, not for embryogenesis or developmental progression (Gambetta and Furlong 2018). The CTCF plays an essential role in the body segment-specific regulation of a particular *Hox* gene, *Abdominal-B* (Gambetta and Furlong 2018).

At minute scale, the DNA is wrapped (147bp) in structures called nucleosomes which consist of an octamer of proteins called histones (two units of each H3, H2A, H2B, and H4). The histones comprise domains of specific amino acid residues which can be reversibly and covalently modified, by adding or removing compounds. (e.g., methyl, phosphate, and acetyl groups) by special type of enzymes such as kinase, phosphatase or methylase/demethylase (Tsankova et al. 2007). Many amino acid residues can be modified in the same tail, these modifications are designated as histone marks and have a specific notation. For example, if the lysine at position 24 of the histone 3 is methylated, this is represented as H3K24me. The histone marks can be recognised by chromatin remodelling proteins, which can further introduce the local changes in the DNA, for example the nucleosome compaction, consequently the DNA can be open (decompacted) and easily accessible to the various regulatory proteins like TFs or RNAP, or the DNA can be closed (compacted), silencing the gene expression (Plass et al. 2013). The histone marks are generally associated with the transcriptional states (active or inactive genes) or cis-regulatory elements (Lawrence, Daujat, and Schneider 2016). For example, H3k27ac is mainly linked with active promoters and distal regulatory elements, H3k4me3 and H3K36me3 are associated with the transcribed chromatin, H3K36me3 is found along the gene body of transcribed genes. By contrast to these active marks, H3K9me3, H3K27me3, and H4K20me3 are usually related to gene repression (Barski et al. 2007).

Overall, the gene expression regulated by the chromatin structure, the various histone marks and DNA methylation are known as epigenomic regulation of gene expression. The transcription regulation by these very factors is a result of the local DNA structure and covalent modifications on DNA and these are the key players that limit or ease the recruitment of regulatory protein molecules.

1.3.2. Transcription Factors

The transcription factors (TFs) basically are the regulatory proteins responsible for activation or deactivation of the gene expression. The main particularity of these TFs are that they recognize very short DNA sequences, known as Transcription Factor Binding Sites (TFBS), varying in length from 6-20 base pairs (bp). The TF binding to regulatory sequences like promoters or enhancers is a fundamental step in initiation of transcription and alteration of transcriptional rates (Nikolov and Burley 1997, Busby and Ebright 1999, Zenkin and Yuzenkova 2015, Hillen et al. 2017). Typically, TFs are mainly categorized as activators of repressors. However, some TFs can act as both activators and repressors depending on the condition (Lee, Minchin, and Busby 2012). To provide a specific example, nuclear receptors (NRs) are ligand-dependent TFs that regulate the gene expression programs in response to small molecules (Huang, Chandra, and Rastinejad 2010). NR transactivation domains (also known as activating domain, AF) have capability to alter the transcription in a context-dependent manner by recruiting various types of multi-protein co-regulatory complexes, often known as co-activators and co-repressors (Huang, Chandra, and Rastinejad 2010). For example, the AF1 and AF2 terminal regions of human glucocorticoid receptor alpha (hGRα) can link receptor to different complexes depending on the cellular signal. In the presence of glucocorticoids, both AF domains interact with transcription-activating factors including basal TFs, co-activators and chromatin modulators. On the other hand, when bound to different gene regulatory regions, these same AFs can, in response to glucocorticoid signals, recruit transcription repression complexes including histone deacetylases (HDACs), chromatin remodelers and down regulate the transcriptional activity of target genes (Huang, Chandra, and Rastinejad 2010).

The feature that differentiates a TF from other regulatory elements, is the capability to bind DNA through a DNA-binding domain (DBD) whereas the other regulatory elements lack the DBD are typically considered as co-factors. The DBDs can recognize major or minor groove of DNA and further create short-term weak interaction between the amino acids of DBDs and nucleotide of the TFBSs. For example, the cell cycle in mammals is controlled by E2F family of transcription factors. With the related dimerization partner

(DP) proteins, E2Fs bind to DNA as heterodimers, whereas E2Fs, E2F7 and E2F8 contain DBDs and act as repressors. The E2F transcription factor family is divided into two subfamilies: E2Fs 1-3 are transcription activators and E2Fs 4-8 are repressors. E2F proteins 1-6 bind to DNA as heterodimer with DP proteins and recently discovered members of the family E2F7 and E2F8 are “atypical” because they possess two distinct DNA-binding subdomains (Morgunova et al. 2015).

Each TF typically recognizes a collection of similar DNA sequences and using *in silico* analysis, we can deduce the consensus sequence of binding (i.e. a representation of the collection of sequences bound by query TF) for a huge number of TFs (Khan et al. 2018, Wasserman and Sandelin 2004, Wingender et al. 1996, Fornes et al. 2019). For *Drosophila*, Redfly is a commonly used database for identification of TFBS for known TFs (Gallo et al. 2011). The identification of TFBSs is a complex task either at computational and experimental level, and it is not completely understood how these TFs recognize their binding sites in the genome. Numerous models have proposed that TF spent a lot of time on DNA searching for their binding sites and by four different modes of motion: (i) 3D diffusion (i.e. the TF moves freely in the nucleus), (ii) 1D sliding (i.e. the TF moves through short regions of DNA), (iii) intersegmental transfer (i.e. the TF moves from one DNA segment to another that are not linearly close) and (iv) hopping (i.e., the TF make short ‘jump’ away from the DNA) (Schmidt et al. 2014, Bauer and Metzler 2012). With these movements TFs can scan hundreds or thousands of nucleotide in very little time and few studies suggest that TFs bind to *bona fide* binding sites that are located in the regions with a similar GC-content to the consensus binding sequence (Slattery et al. 2014, Dror, Rohs, and Mandel-Gutfreund 2016). It is widely known that within TFBSs not all the nucleotides contribute with the same strength for the TF binding, usually the strongest nucleotide contributors are the most conserved position in the consensus, however many studies showed that the flanking sites are determinants of strong or weak TF binding specificity (Schöne et al. 2016). For example, the glucocorticoid receptor (GR), a member of the steroid hormone receptor family, binds as homodimer to genomic response elements, which have particular sequence (Schöne et al. 2016). The nucleotides flanking the core-binding site, differ depending on the strength of GR-dependent activation of nearby genes (Schöne et al. 2016). Schöne et al. also found that these flanking nucleotides change the 3-D structure of the DNA binding site, the DNA-binding domain of GR and the quaternary structure of the dimeric complex (Schöne et al. 2016). Genes regulated by strong promoters yeild more mRNAs and therefore more product protein than genes regulated by weak promoters. This is useful because some proteins are required in abundance whilst others are required in low quantities.

1.4. Gene expression during *Drosophila* spermatogenesis

Developing spermatogonia can dedifferentiate and re-gain stem cell properties and contribute to the stem cell population. Differentiation of spermatogonia into primary spermatocyte can be called as a dramatic stage in *Drosophila* germline cells because during this transition primary spermatocytes activate a specialized transcriptional programme. Comparative gene expression analysis in different adult fly tissues revealed that in genome ~50% of the genes are expressed in testes and 8% of the detected transcripts in adults are testis- specific, while a further 5% are testis-enriched (Andrews et al. 2000, Chintapalli, Wang, and Dow 2007).

To analyse the specific stages of spermatogenesis when many of these transcripts are expressed, a large scale RNA *in situ* hybridisation project analysed 553 transcripts and 529 transcripts were found to be transcribed in primary spermatocytes, persisted in spermatid cytoplasm and were degraded at different stages in elongation (Barreau et al. 2008). Genes expressed in spermatocytes, for example, *male sterile (3)K81* (encodes a telomere capping protein) show high levels of expression and exclusively expressed in primary spermatocytes but it shows an effect after fertilization (Gao et al. 2011). In male germ line, K81 replaces HipHop in spermatid nuclei (Dubruille et al. 2010) and at fertilisation, K81 transmits to the zygote, where it protects paternal telomeres and prevent end-to-end fusion of different chromosomes during the first division. If this process fails then the chromosomes in the male pronucleus end up fused, and thus not able to do mitosis properly after fertilisation. Although K81 is required for telomere protection at fertilisation, it is not required during spermatogenesis (Dubruille et al. 2010, Gao et al. 2011). In addition to above, genes such as *schuy* and *CG31858* show low levels of expression in mitotic phase followed by an increase in expression during the meiotic phase and high levels of expression in the post meiotic phase (Vibranovski et al. 2009, Barreau et al. 2008).

Newly discovered segregating as well as fixed *de novo* genes (discussed in section 1.5.4) are mostly expressed in late spermatogonia and early spermatocytes and spermatids express least *de novo* genes (Witt et al. 2019). The high proportion of *de novo* genes expressed in spermatocytes suggest that these genes could play vital roles in spermatogenesis (Witt et al. 2019).

1.4.1. Activation of specialized transcriptional programme in primary spermatocytes

The spermatogonia undergoes four mitotic divisions before differentiating to primary spermatocytes. For terminal differentiation, primary spermatocytes exhibit high levels of transcription. Different male sterile mutants have been identified at relatively few different arrest points during spermatogenesis. General classes are– no stem cells, spermatogonia over proliferation, spermatocyte arrest, and no individualisation (Wakimoto, Lindsley, and Herrera 2004). These points include a group of genes that, when mutated, cause arrest at the end of primary spermatocyte development **(Figure 1.2)**. These genes are widely known as meiotic arrest genes because mutants arrest at the G2-M transition of meiosis-I (Lin et al. 1996). Meiotic arrest genes are essential for meiotic cell cycle progression as well as the onset of spermatid differentiation into mature spermatocytes. Initially meiotic arrest genes were subdivided into two classes (White-Cooper et al. 1998) but some of the more recently identified mutants do not fit into either category - *Nxt1*, *mediator*, *magellan*, *Mi-2, thoc5* (Lu, Kim, and Fuller 2013, Boube et al. 2014, Caporilli et al. 2013, Murawska et al. 2008, Kunert et al. 2009). The *aly* class genes, among which *always early* (*aly)* and *cookie monster* (*comr)* **(Table 2)***,* regulate the transcription of few genes required for entry into meiosis (*boule, twine, Cyclin B*) also, many other genes required for the differentiation of functional sperm (*fuzzy onions, janus B, don jaun, gonadal*). Other class, the *can* class genes *cannonball* (*can*),*meiosis I arrest* (*mia*)*, and spermatocyte arrest* (*sa)* **(Table 2)** regulate meiotic cell cycle indirectly by regulating a factor that controls translation of *twine* but are required for transcription of genes important in spermiogenesis (White-Cooper et al. 1998).

Lin *et al.* 1996 (Lin et al. 1996) found out that in *can*, *mia* and *sa* mutants the chromosomes in the arrested cells are partially condensed (like prophase I). In addition to above, in *aly* mutant spermatocytes, the chromatin morphologically was found to be disrupted and chromosomes were fuzzy (Lin et al. 1996). White-Cooper *et al.* 1998 (White-Cooper et al. 1998) found out that genes required for the functioning of primary spermatocytes were expressed in mutants of meiotic arrest genes suggesting mutant arrest genes are not responsible for global activation of the transcription in primary spermatocytes.

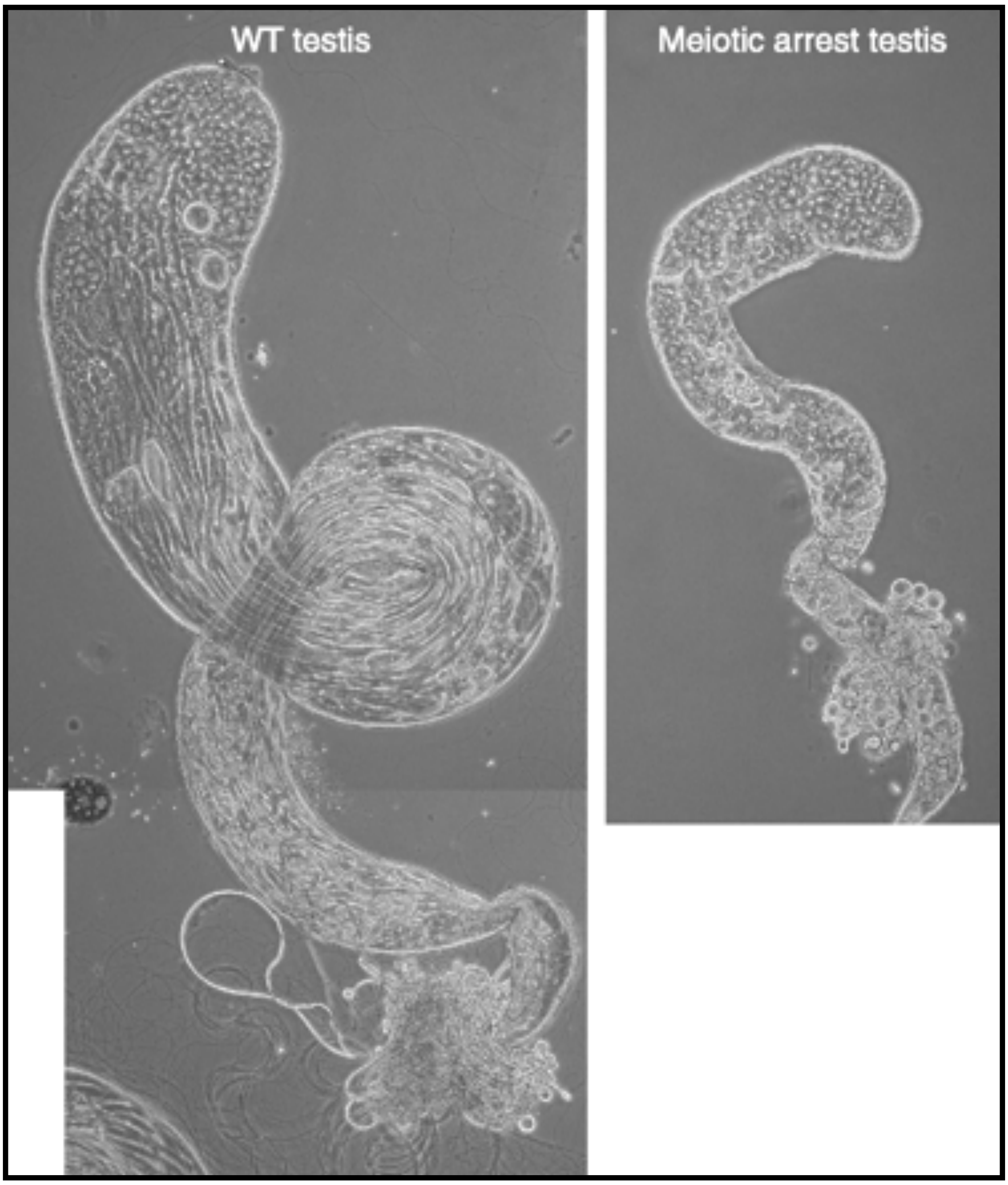


Figure 1.2:- Wild type and meiotic arrest mutant testes (*can* class- *sa*) of *Drosophila* (Figure adapted from White-Cooper *et al.* 2010).

Genes required for spermiogenesis were reduced in expression in all meiotic arrest mutants, although more dramatically in *aly* mutants than in *can, mia* or *sa* and suggest that meiotic arrest genes are important for gene expression regulation in primary spermatocytes (White-Cooper et al. 1998, Doggett et al. 2011). More than 1000 genes are downregulated in *aly*-class mutant whilst a restricted set of genes are affected in *can*-class mutants (Doggett et al. 2011, Caporilli et al. 2013, Chen et al. 2005, Laktionov et al. 2014, Hiller et al. 2004)

Gene symbol	Gene full name	Classification	Molecular Function
can	*cannonball*	*can*-class (Hiller, Lin et al. 2001)	TAF5
mia	*meiosis-I arrest*	*can*-class (Hiller, Chen et al. 2004)	TAF6
sa	*spermatocyte arrest*	*can*-class (White-Cooper et al. 1998)	TAF8
nht	*no hitter*	*can*-class (Hiller, Chen et al. 2004)	TAF4
rye	*ryan express*	*can*-class (Hiller, Chen et al. 2004), (Metcalf and Wassarman 2007)	TAF12
mip40	*Myb interacting protein, 40kDa*	*can*-class (Beall, Lewis et al. 2007)	Component of dREAM complex, regulating transcription of cell cycle and developmental genes
aly	*always early*	*aly*-class (White-Cooper, Leroy et al. 2000)	*lin-9* homologue, regulation of transcription of testis genes
comr	*cookie monster*	*aly*-class (Jiang and White-Cooper 2003)	DNA binding
tomb	*tombola*	*aly*-class (Jiang, Benson et al. 2007)	DNA binding
topi	*matotopetli*	*aly*-class (Perezgasga et al. 2004)	DNA binding
achi-vis	*achintya and vismay*	*aly*-class (Ayyar, Jiang et al. 2003), (Wang and Mann 2003)	DNA binding

caf1/p55	*Chromatin assembly factor 1, p55 subunit*	*aly*-class (Wen, Quan, and Xi 2012)	WD repeat domain protein homologous to human Rb associated proteins (RbAp48, RbAp46). Component of dREAM complex, regulating transcription of cell cycle and developmental genes
wuc	*wake-up-call*	No mutant alleles (Doggett et al. 2011)	Unknown

Table 2:- *Drosophila* Meiotic arrest loci, identity, classification, and molecular function.

Five other meiotic arrest genes were identified and characterised, which cannot be classified as either *aly* or *can* class. Wake-up-call (Wuc) was identified by yeast-2-hybrid screen and it shows physical interaction with Aly (Jiang et al. 2007). Wuc is highly expressed in primary spermatocytes and associated with chromatin. However, unlike testis-specific TATA binding protein-associated factors (tTAF) or testis-specific meiotic arrest complex (tMAC) mutants, loss of *wuc* does not perturb the expression of meiotic cycle genes or spermatids differentiation (Doggett et al. 2011). Another study revealed that disruption of a component of the THO complex, THOC5, led to meiotic arrest phenotype. This complex is known for exportation of mRNAs from nucleus to cytoplasm. Loss of *thoc5* does not abolish the mRNAs export or affect the meiotic cell cycle genes or spermatid differentiation genes (Moon et al. 2011). A study identified Nxt1 also a component of mRNA transport machinery, required for accumulation of mRNAs (Caporilli et al. 2013). The dependence of these mRNAs on Nxt1 has a distinct mode compared to tTAF or tMAC-dependent genes (Caporilli et al. 2013). Additionally, during characterization of a meiotic arrest mutant *magellan (magn),* the Magn protein has been shown to regulate the spermatocyte chromatin structure, meiotic cell cycle, and spermiogenesis. In spermatocytes, Ubi-p63E acts in protein degradation-independent manner and it appears that the level of free ubiquitin is critical for normal spermatogenesis progression (Lu, Kim, and Fuller 2013). Lastly, a study identified a novel meiotic arrest gene *kumgnag (kmg)*, which is a Zn-finger containing protein. The expression of *kmg* is turned on in early spermatocytes, independent of tTAF or tMAC. Kmg is required for maintenance of germline identity by suppressing the expression of hundreds of somatic cell genes whose expression is normally found in somatic cells. Kmg also acts with chromatin remodeler dMi-2 and restrict the tMAC component Aly from helping to fire transcription of somatic genes by repression of cryptic promoters (Kim et al. 2017).

1.4.2. *cannonball* (*can*)-class

Currently all the characterized *can*-class genes encode proteins known as tTAFs. TAFs and TATA binding protein (TBP) form a complex called Transcription Factor II D (TFIID) which is a critical component of basal transcriptional machinery that further recruits and assists RNA polymerase II (RNApol II)**.** To prepare for transcription, a complete set of transcription factors and RNA polymerase need to be assembled at the core promoter (He et al. 2013, Lenhard, Sandelin A Fau - Carninci, and Carninci , Lenhard, Sandelin, and Carninci 2012). TFIID initiates the assembly of a transcription complex and other general transcription factors stabilizes DNA-TFIID complex and allow recruitment of RNApol II in association with other general transcription factors (Hendrix et al. 2008, El-Sharnouby, Redhouse, and White 2013, Jiang et al. 2018, Kurshakova et al. , Li et al. 2009)

From the cloning experiment of *can*, it was found out it encodes for a testis specific paralogue of TAF and dTAF5 (dTAFII80) (Hiller et al. 2001). Since, *mia, sa, rye, and nht* also encode tissue specific TAFs, this led to a hypothesis about the presence of testis specific TFIID complex. *mia* encodes TAF6; TATA Binding Protein (TBP) associated Factor 6 (TAF6) is one of the several factors that bind TBP and is involved in forming Transcription Factor IID (TFIID). It is also a component of histone acetyl transferase (SAGA). *sa* encodes a TAF8; TBP Associated Factor 8 and also possess bromodomain. *rye* encodes for a TFIID; Transcription initiation factor TFIID subunit A (TAF12) and *nht* encodes TAF4; TBP Associated Factor 4 (TAF4) is one the several TAFs that bind TBP and is involved in forming Transcription Factor IID (TFIID) complex. The testis specific TAFs (tTAFs) which are encoded by *can*-class meiotic arrest loci later complexes with a TAF1-2 (TAF1-2 is a testis-specific splice isoform of TAF1) (Chen et al. 2005, Metcalf and Wassarman 2007). TAF1 is Bromodomain containing TFIID-like subfamily, the largest subunit of TFIID complex and initiates the assembly of the transcription machinery.

Many TAFs are histone modifiers implicating a role in regulating transcription through altering the chromatin confirmation and allowing transcriptional machinery to access it (Grant, Schieltz et al. 1998, Ogryzko, Kotani et al. 1998, (Denslow and Wade 2007). TAFs are ubiquitously expressed and are known to form complexes with the repressive polycomb group (PcG) proteins in *Drosophila* embryos (Saurin, Shao et al. 2001) and tTAFs have been exhibited to co-localize with PcG proteins (Polycomb, Polyhomeotic and dRING proteins) in *Drosophila* spermatocytes (Chen, Hiller et al. 2005). tTAFs have been hypothesized to allow the transcription of target genes by counteracting the repression caused by PcG proteins, perhaps by causing the dissociation of Polycomb repressive complex I (PRC 1) from cis-regulatory sequences. tTAFs are found concentrated with the nucleolus of primary spermatocytes (Chen Hiller et al. 2005). Contrary studies using Chromatin Immunoprecipitation (ChIP) showed that there was no enhancement of Polycomb at promoters of tTAF dependent genes in tTAF mutant compared to wild type (El-Sharnouby, Redhouse, and White 2013). The absence of Polycomb at tTAF-dependent spermatogenesis genes argues against a model where Polycomb displacement is the mechanism of gene activation during spermatogenesis (El-Sharnouby, Redhouse, and White 2013).

1.4.3. *always early* (*aly*)-class

aly was cloned, and found to be homologous to *C. elegans lin-9* (White-Cooper et al. 2000). A second *Drosophila* homologue of *lin-9* is present in the genome- *Mip130*. Mip130 was purified as component of a complex containing *Drosophila* Myb (Myb interacting protein 130) (Beall et al. 2002). Alternative purification strategies revealed related complexes - dREAM and MMB, which contain Myb, CAF1, Mip120, and Mip40 in addition to mip130. Under different conditions, this complex has been re-purified and revealed the presence of more subunits of this complex, it is known as the Myb-MuvB (MMB) or dREAM complex (Korenjak et al. 2004), (Lewis et al. 2004). Additional subunits of this complex contain Rbf (homolog of Retinoblastoma (Rb), E2F2, Dp and dLin52. dREAM/MMB complex is related *Drosophila* tMAC, as it contains two of the same components, Mip40 and Caf1. A further two proteins, Mip120 and Mip130 show significant sequence similarity to tMAC proteins Tomb and Aly respectively. The dREAM complex comprises multiple DNA binding proteins known to act as both activators and repressors of transcription. Components Myb, E2F/DP and Mip120 have site specific DNA binding ability and Mip130 contains an A-T hook domain which is predicted to be capable of binding to AT-rich DNA sequences.

In addition to MMB complex, tMAC a testis specific meiotic arrest complex was identified using combination of different chromatographic techniques including affinity chromatography, ion exchange chromatography, and gel filtration chromatography (Beall et al. 2007). During this study, the only MMB subunits found were Mip40 and Caf1/p55, in addition to other testis specific proteins Aly, Comr, Topi and Tomb (Beall et al. 2007). When *comr* was cloned, it was found that Comr contains a winged helix putative DNA binding domain (Perezgasga et al. 2004). *topi* was cloned based on its direct interaction with *comr* protein, and contains multiple Zn-finger motifs and putative DNA binding site (Perezgasga et al. 2004). When *tomb* was cloned through a screening of *aly* protein binding partners, it was found to contain a CXC predicted DNA binding domain, and is paralogous to the dREAM component Mip120 (Jiang et al. 2007).

Another *aly*-class meiotic arrest locus identified using genetics was *achi*+*vis* gene duplication locus, encode Achi and Vis proteins. Achi and Vis are almost identical proteins related to human transforming growth-interacting factor (TGIF) (Ayyar et al. 2003, Wang and Mann 2003). Achi /Vis proteins co-immunoprecipitated with Aly and Comr from testis extracts but do not co-purify with Mip40 (Perezgasga et al. 2004), suggesting that Aly and Comr are in at least two different complexes, one with Mip40 (Laktionov et al. 2018), but lacking Achi/Vis, other containing Achi/Vis. The following table represents homologous complexes in other organisms compared to a tMAC complex of *Drosophila.*

H. Sapiens **DREAM/LINC**	***C. elegans*** **MuvB/DRM (vulva development)**	***D. melanogaster*** **dREAM/Myb-MuvB**	***D. melanogaster*** **tMAC**
RBL2/p130 (testis)	Lin-35	RBF1 or RBF2 (testis)	
E2F4 or E2F5 (testis)	Efl-1	E2F2 (testis)	
DP1 (testis)	Dpl-1	DP (testis)	
RBBP4 (testis)	Lin-53	Caf/p55 (testis)	Caf1/p55 (testis)
LIN9 (testis)	Lin-9	Mip130(testis)	Aly (testis)
LIN37 (testis)	Lin-37	Mip40(testis)	Mip40 (testis)
LIN52 (testis)	Lin-52	dLin52(testis)	Wuc (testis)
LIN54 (testis)	Lin-54	Mip120 (testis)	Tomb (testis)
MYB-B/MYBL2 (testis)		Myb (testis)	
			Comr (testis)
			Topi (testis)
			Achi/Vis (testis)

Table 3:- Comparison of homologous complexes in different organisms compared to *Drosophila* testis specific complex-tMAC (Table adapted from White-Cooper *et al.* 2010).

Yeast 2-hybrid studies carried out by Dogget., et al. found out CG12442 interacts with Aly and later this gene named as a wake-up-call (wuc) (Doggett et al. 2011). In addition to Aly, Wuc also interacts with Comr and Topi and it was shown by pair-wise yeast 2-hybrid experiments (Doggett et al. 2011).

Furthermore, *comr* (*aly* class gene) plays a vital role to promote transcription and allow primary spermatocytes to initiate meiosis and later spermatid differentiation. (Jiang and White-Cooper 2003). *comr* gene encodes for a nuclear acidic protein with predicted DNA binding domain (Jiang and White-Cooper 2003). The expression level of *comr* changes during the development of spermatocytes. In early spermatocytes, *comr* has a high level of mRNA expression but as spermatocytes grow, the levels of *comr* decreases and in primary spermatocytes it is undetectable (Jiang and White-Cooper 2003). Expression of *comr* is independent of either *aly* or *can* classes.

1.4.4. Model of *aly*-class meiotic arrest gene function

It is supposed that the Achi/Vis, Topi and Tomb proteins bind to specific promoter sequences of testis specific genes. The Achi/Vis, Topi and Tomb proteins are not required for nuclear localisation of Aly or Comr but are required for localisation to chromatin (Laktionov et al. 2014). A proposed model suggests that Achi/Vis, Topi and Tomb bind to specific promoter regions and then recruit Aly and Comr to form a complex **(Figure 1.3)**. In comr mutant Aly protein stays in the cytoplasm, and when aly is mutant Comr protein stays in the cytoplasm therefore it is likely that these two proteins need to interact directly or indirectly to maintain each other’s nuclear localisation **(Figure 1.3)**.

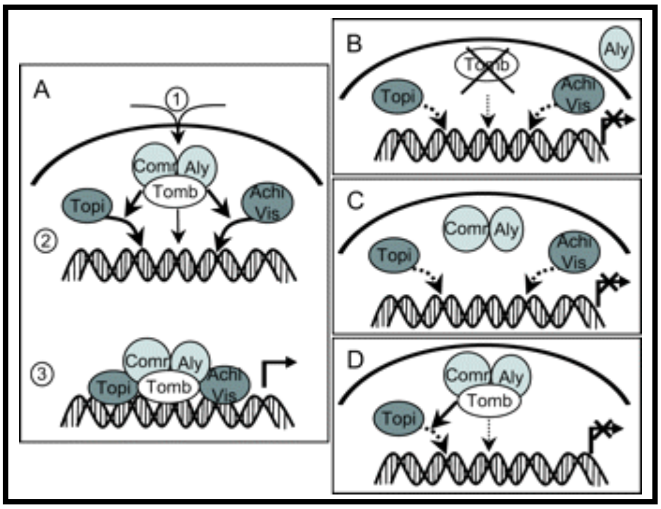


Figure 1.3:- Model of assembly of *aly*-class meiotic arrest proteins at target promoters (taken from Jiang, Benson et al. 2007).

A large set of genes are dependent on the aly-class meiotic arrest genes for their transcription. Microarray has identified possibly 1200 transcripts are dependent on aly-class in primary spermatocytes (Barreau et al. 2008, Jiang and White-Cooper 2003, White-Cooper et al. 2000, Hiller et al. 2001). aly-class meiotic arrest genes regulate the genes whose product are likely to function in late in spermatogenesis.

1.4.5. Testis meiotic arrest complex (tMAC)

tMAC is a testis-specific paralog of MMB (Beall et al. 2007). This complex was named as tMAC complex, because *aly, comr, topi*, and *mip40* mutants all show the same testis phenotype, i.e an arrest at the primary spermatocyte stage of development. These components work together in a complex and promote differentiation and meiotic cell cycle progression. Other proteins might interact with tMAC to help in the regulation of testis-specific transcripts. Both the tMAC **(Figure 1.4)** and MMB contain proteins, which are similar to each other in domain architecture except Mip40 and Caf/p55: Aly (tMAC) or Mip130 (MMB) and Tomb (tMAC) or Mip120(MMB). The tMAC and MMB contain multiple site-specific DNA-binding proteins- Topi and Tomb in tMAC and Myb, E2F2/DP, Mip120, Mip130 in MMB.

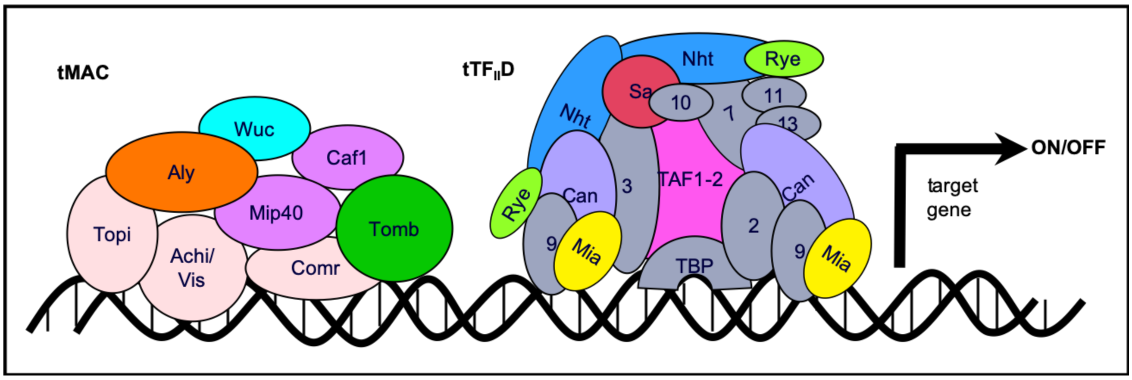


Figure 1.4:- Model of assembly of tMAC and $tTF_{II}D$ complex at target promoter.

In gonads, components of the RNA polymerase II transcription machinery are crucial for developmentally regulated gene expression in testis. Testis specific TATA binding protein-associated factors (TAFs) are encoded by the can class genes- *can, sa, mia nht,* and *rye* genes (Hiller et al. 2001). These TAFs creates a testis specific TFIID (tTFIID) complex that drives gene expression of developmentally essential genes in spermatocytes **(Figure 1.4)**. The *mip40*-null allele is also in the can class, suggesting tMAC might be interacting with tTFIID at many of promoters and all the genes that require tTAFs for expression also require tMAC (Beall et al. 2007). Contrary data by HW-C (unpublished data) showed that the allele of *mip40* in Beall et al. 2007 is not null but it is still a can class.

1.4.6. Testis-specific promoter

Testis-specific genes are activated only in testes and kept silent in other tissues of the fly. There is relatively little known about the promoter elements conferring testis specificity. Testis-specific β-tubulin isoform β2tubulin (βTub85D) was the first to be studied. Astoundingly, a fragment comprising of only 53 bp of promoter region, plus the first 71 bp of the 5'UTR was enough to confer testis-specific expression on reporter gene. Within promoter, motif of 14 bp, β2UE1 was identified to be critical for testis-specific expression (Michiels et al. 1989). The β2UE1 motif sequence cannot be considered as signature sequence because this was found upstream of several testis-specific transcriptional start sites (TSS) although this sequence is not present in most of other testis promoters, so cannot be considered a signature sequence for testis-specific expression (Nurminsky et al. 1998, Yang, Porter, and Rawls 1995). These short promoters most probably contain a landing site for the transcriptional machinery, tTAFs, and tMAC (as discussed above).

Another study identified a core promoter sequences of 190 genes specifically expressed in testis that have a 10 bp A/T-rich motif that is identical to the translational control element (TCE). Mutation of TCE reduced the transcription of reporter gene indicating that the TCE is important but not essential for activation of transcription. Moreover, the TCE regulates transcription both dependently and independently of testis specific TFIID (tTFIID) (Katzenberger et al. 2012). Testis-specific control elements are so small and may be important in allowing new gene duplicates to be expressed in testes. For example, *Sdic*, a testis-specific gene evolved from duplication and fusion of the two genes *AnnX* (encoding annexin X) and *Cdic* (cytoplasmic dynein) (Ranz et al. 2003). The fusion joins *AnnX* exon 4 with *Cdic* intron 3, which brings together three putative promoter elements for testes-specific expression of Sdic: the distal conserved element (DCE) and testes-specific elements (TSE) are derived from *AnnX*, and the proximal conserved elements (PCE) from *Cdic* intron 3. *Sdic* transcription initiates within the PCE, and translation in initiated within the sequence derived from *Cdic* intron 3 (Ranz et al. 2003).

1.4.7. Translational delay

Extensive cell growth, a high level of mRNA, and protein synthesis are the characteristics of the prolonged primary spermatocyte developmental stage. During this stage of development, all the studied genes expressed exclusively at this stage have short and strong regulatory regions (Blümer et al. 2002, Katzenberger et al. 2012). In Drosophila, sperm morphogenesis takes 3.5 days and dormant mRNAs need to be recruited for translation at distinct times in order to coordinate the expression of certain proteins with respect to the changing requirements of the cell during sperm assembly (Blümer et al. 2002). Consequently, many mRNAs are translationally repressed and get activated for translation during spermatid morphogenesis. Hence, translational control is a vital characteristic of spermatogenesis. Several genes for example, *Mst(3)CGP, dhod, janus B* and *don juan (dj)* encode translationally repressed mRNAs. *Mst(3)CGP* has been an extensively studied gene and has 12 bp sequence motif in the 5'UTR called TCE (translational control element). The TCE is important for translational repression of the *Mst(3)CGP* mRNA in meiotic stages. Onset of translation is delayed until fully elongated spermatid stage and secondary polyadenylation aids the activation of translation (Kuhn et al. 1991). Sequence similarities to the TCE were found in a number of translationally controlled mRNAs and suggest that transcripts of the many genes are regulated by the same regulatory element (Blümer et al. 2002). The TCE-like motif containing regions do not always participate in the translation repression as well, such as *dj* that utilizes a distinctive translational repression elements. The TCE can act as translational repression or activation element in testis specific genes (Blümer et al. 2002).

1.5. Origin of *de novo* genes

Recent advances in high throughput transcriptome analyses has been helping scientists to study gene expression regulation in depth. Such advancement addressed many scientific mysteries and changed the way of thinking of researchers about central dogma of life and opened many new adventurous journeys in science. Finding of the *de novo* genes is one of the examples of previously unexpected phenomena being discovered through the application of high-throughput sequencing and bioinformatics. Most new genes are derived from various different mechanisms like gene duplication, retroposition, genomic rearrangement whereas *de novo* genes are derived from the ancestral non-genic sequence (Hughes 2005, Betrán, Thornton, and Long 2002, Long et al. 2013). The process that govern *de novo* gene birth are not well understood. The first evidence for *de novo* genes derived from studies of protein-coding genes specific to the *D.melanogastor* and *D. yakuba* lineages respectively (Levine et al. 2006, Begun et al. 2006, Begun et al. 2007). Reinhardt et al, 2013 proposed two possible models explaining the origin of *de novo* genes. New protein coding gene evolved from the noncoding sequence. For a non-genic sequence to become a protein-coding gene, it must have to evolve transcriptionally as well as with protein coding potential (Reinhardt et al. 2013).

1.5.1. Proposed models suggesting the origin of de novo genes

De novo genes are defined as having emerged from non-coding DNA sequence must be transcribed and acquire an ORF before becoming translated (Kaessmann 2010). The transcription first model suggest that protein-coding *de novo* genes may first exist as RNA gene intermediates. The case of bifunctional RNAs, which are both translated and function as RNA genes, shows that such a mechanism is plausible. In the ORF-first model, Proto-open reading frames acquire mutations such as, point mutation, insertion or deletion that causes abolishment of frame disruptions and thus generation of a long ORF (Kaessmann 2010). Transcriptional activation of these ORFs happens through the acquisition of promoter located in the 5' flanking region. After acquisition of both an ORF and transcriptional activation of the *de novo* gene is able to produce a protein product, which potentially has a function, and thus a fitness effect. Selection can then act on the activity of the gene and it is possible for such a gene to become essential for viability or fertility (Kaessmann 2010). For birth of a *de novo* protein coding gene to happen, a non-coding sequence must both transcribe and acquire an ORF before being translated. These events can happen in theory in either order, and these is an evidence supporting both an “ORF” first and a “transcription first” model (Schlötterer 2015). Analysis of *de novo* genes that are segregating in *D. melanogaster* with respect to their expression found that sequences that are transcribed have similar coding potential to the orthologous sequences from lines lacking evidence of transcription (Zhao et al. 2014), supporting the notion that many ORFs, at least, exist prior to being transcribed.

1.5.2. Structural features of de novo genes

Normally a gene has different types of features like promoter, regulatory sequences, open reading frames, several exons, number of alternative transcripts, and the absence of satellite sequences (repetitive DNA sequences). Likely, *de novo* genes also contain these features which are given below:

a) Transcript length:- In *de novo* genes, the transcript are usually shorter in length (~801 bp) when compared with the other annotated genes (~1557 bp) (Palmieri 2014, Zhao et al. 2014, Neme and Tautz 2013).

b) No of Exons:- Usually, old genes possess greater number of exons in their body (2.37) whereas *de novo* genes tend to have fewer exons (1.47) than old genes (Neme and Tautz 2013, Palmieri, Kosiol, and Schlötterer 2014).

c) Microsatellites:- *De novo* genes are enriched in microsatellite sequences (Toll-Riera et al. 2008, Palmieri 2014, Wj 2007).

d) Expression levels:- The expression level of *de novo* genes (7.78 FPKM) is usually lower than male-specific annotated genes (66.54 FPKM), studied by RNA sequencing using testes (Donoghue et al. 2011, Palmieri, Kosiol, and Schlötterer 2014, Zhao et al. 2014).

f) Tissue specificity:- The *de novo* genes expression is more tissue specific than old genes (Donoghue et al. 2011, Toll-Riera et al. 2008, Levine et al. 2006).

g) Chromosomal location:- In most cases, *de novo* genes are enriched on the sex chromosome i.e. X-chromosome but in *Drosophila de novo* genes are enriched on chromosome 2 and 3 (Zhao et al. 2014, Levine et al. 2006, Palmieri, Kosiol, and Schlötterer 2014).

1.5.3. *de novo* genes in other organisms

A combination of phylogenetic, genomic/transcriptomic analyses revealed evidence of lineage-or species-specific *de novo* transcripts associated with non-genic orthologous sequences in sister species (Zhao et al., 2014). In addition to *Drosophila, de novo* genes were also found in humans (e.g. *XLOC_196865*), mouse (e.g. *8030423J24Rik*), and yeast (e.g. *YPR126C*), *C. elegans (K09F5.4)* (Zhou et al. 2008, Knowles and McLysaght 2009, Xiao et al. 2009, Cai et al. 2008, Carvunis et al. 2012, Heinen et al. 2009, Zhang et al. 2019) etc. Bioinformatics analysis found out 634 human-specific genes (1,029 transcripts), 780 chimpanzee-specific genes (1,307 transcripts). Taken together, the total number of candidate *de novo* genes was 1414 (2336 transcripts) (Ruiz-Orera et al. 2015). The transcripts of *de novo* genes in human and chimpanzee are enriched in testis (93.8-94.5%) (Ruiz-Orera et al. 2015). Ruiz-Orera et al, 2015 also found *de novo* genes were underrepresented in other organs like the brain, liver, and heart when compared to other transcripts.

1.5.4. *de novo* gene expression in *Drosophila* testis

Zhao et al, 2014 used Illumina paired-end RNA sequencing to characterize the transcriptome of six highly inbred lines established by the DGRP (*Drosophila* Genetic Reference Panel, lines RAL-304, RAL-307, RAL357, RAL360, RAL-399 and RAL-517) *D. melanogaster* strains. A total 142 segregating and 106 fixed testis specific *de novo* genes were found to be expressed in *Drosophila* testis (Zhao et al. 2014).

Segregating de novo genes:- These genes were not annotated in the reference annotation, were not expressed in other tissues, and orthologous regions were also not transcribed in closely related *Drosophila* species (*D. simulans and D. yakuba*). Thus these represent testis-specifically expressed *de novo* genes with both the derived, expressing, alleles and the ancestral, non-expressing alleles, still segregating in the wild *D. melanogaster* population.

Fixed de novo genes:- These genes are unannotated male-specific expressed in all six DGRP strain and in reference strain (*D. simulans*) but not in outgroup strain (*D. yakuba* strains- Tai18E2 and CY28) thus it represents testis-specifically expressed *de novo* genes are fixed in the wild *D. melanogaster* population

To identify segregating *de novo* genes, the authors used the RNA-sequencing data from each strain were used for *de novo* transcriptome assembly and a reference sequence guided transcriptome assembly (Zhao et al. 2014). These two approaches are complementary, and as expected, were highly concordant. The authors used the assemblies to determine which of the well-supported transcripts were not annotated in the reference sequence or supported by *Drosophila* modENCODE data and thus, were potentially segregating *de novo* genes. To increase the likelihood that candidate transcripts corresponds to *de novo* genes, authors used several criteria. 1) Putative *de novo* transcripts were required to map to an intergenic regions. 2) Candidate transcripts were required to be located at least 500 bp away from known gene. 3) Minimum transcripts size was 200 bp and the minimum expression required was FPKM>2 (FPKM: fragments per kilobase of exon per million fragments mapped). 4) Splice junctions were required to be covered by at least 20 reads. 5) Intron and exon size > 50bp. 6) No BLASTP hit for the candidate gene in the NCBI nr (nonredundant) protein database (Zhao et al. 2014). The authors repeated this analysis on testes from three inbred strains of *D. simulans* and two *D. yakuba* and used these outgroup data and the parsimony criterion to determine whether polymorphic *D. melanogaster* transcripts was described as a *de novo* gene rather than a loss of expression. Zhao et al., also used *D. simulans* whole female and male RNA-seq population data from 59 isofemale strains to check if *de novo* genes is expressed in other tissues of outgroups and reduce the possibility of a shift of ancestral gene expression confounding the inference of gene gain. All main conclusions from these analysis remain if minimum FPKM estimate for expression is FPKM>1 (Zhao et al. 2014).

RNA sequencing suggests that each *de novo* gene is expressed in some *Drosophila* lines but is not expressed in other lines. This suggests that their expression must depend on factors that are present in testis. Zhao, L. et al., also found that *de novo* genes were significantly shorter in length and simpler than annotated genes and male-biased genes (genes express at higher levels in male than female). 57% genes had a single exon compared to annotated genes as well as male-biased genes 94% (134) of *de novo* genes had a minimum open reading frame (ORF) of at least 150 bp and had coding potential. In putative protein-coding genes the average 5' and 3' untranslated region (5' UTR and 3' UTR) lengths were-248 bp and 364 bp respectively. These were slightly shorter than the average lengths for annotated *D. melanogaster* genes (5'UTR~282 bp and 3' UTR~503 bp) but were slightly longer than the averages for annotated male-biased genes (5'UTR~239 bp and 3' UTR~392 bp). For comparative analysis, male biased genes were used because *de novo* genes tend to be specifically expressed in tissue associated with male reproduction and suggest that sexual or gametic selection may be involved (Zhao et al. 2014).

In *Drosophila, de novo* gene expression is mainly influenced by variation in the cis-acting regions i.e., *de novo* gene expression is due to single nucleotide change present in 5' flanking regions of expressing chromosome (Zhao et al. 2014). Potential regulatory elements in segregating *de novo* genes share four common 8- and 10-bp consensus motifs in 5' UTR and these motifs are mostly found in multiple-exon genes. Male annotated genes share at least one of these motifs and suggests that *de novo* genes share regulatory elements with known male-biased genes. This suggests that de novo gene expression is mainly influenced by cis-acting variants in the regions corresponding to the 5' flanking regions. A single nucleotide change or insertion or deletion in the noncoding region of gene i.e. promoter could regulate the gene expression but this was not tested whether the SNPs had any functional significance (Zhao et al. 2014). Zhao et al, also found, different single-nucleotide polymorphisms among flanking regions of expressing alleles and non-expressing alleles (about 500bp upstream of the TSSs of the polymorphic *de novo* genes).

To determine the overall patterns of gene loss and gain, authors compared segregating and fixed gene gains and losses in the *D. melanogaster* lineage. There were many more gene gains than gene losses (both segregating and fixed). Zhao et al, also found that substantial number of segregating *de novo* gene losses and observed no population genetic evidence that losses result from directional selection. These *de novo* genes may often spread under selection, while gene loss may occur because of drift associated with loss of ancestral gene function (Zhao et al. 2014).

A majority of the *de novo* genes found are segregating *de novo* genes and this is the key feature that made them so useful for our research. Studying these genes could help us understand the mechanism of gene expression regulation during spermatogenesis. It would also tell us the role of tMAC in gene expression regulation of these *de novo* genes. In our lab, we have selected previously uncharacterized alleles of 72 segregating *de novo* genes which are present on 3rd chromosome and we chose these genes because tMAC component- aly is present on 3rd chromosome and it would allow us to understand the role of tMAC in gene expression regulation using genetic tools. We have analysed the expression of these 72 *de novo* genes in phenotypically normal and *aly* meiotic arrest mutant testes. For each gene we determined whether the allele present on an *aly2* mutant chromosome (aly is present on 3rd chromosome) was expressed, and, if expression was detected, whether that expression was dependent on, or independent on *aly*. We found out that 25/63 of the *de novo* genes are tMAC dependent for their expression (for example-Gene 3L_061) and 9 are expressed in the absence of tMAC (for example Gene 3L_062). In our lab, we have tested gene_068 which has few SNPs encompassing the TSS. We designed and cloned synthetic DNA fragments in reporter constructs with combination of SNP. The reporter constructs showed that no SNPs play a role in gene expression regulation.

For my thesis, we chose two segregating *de novo* genes, gene_074 and gene_090. Bioinformatics analysis of these two genes (074 and 090) found that expressed allele and non-expressed allele of each gene possess variation in their promoter region for example insertion, deletion, and single nucleotide polymorphisms (SNPs). This suggests that presence of SNPs or indels in the promoter region of these genes could be critical for turning ON and OFF gene expression.

gene_3L_090: - gene_090 is present on the left arm of 3rd chromosome of *Drosophila* and expressed in RAL307, 360, 399, 517 lines but is not expressed in RAL304 and 357 (Zhao et al. 2014). RNAseq data from modENCODE project shows that gene_090 is only expressed in gonads but not in other developmental stages of *Drosophila* **(Figure 1.5)**. Interestingly, allele of gene_090 from fully sequenced *aly2* chromosome and only expresses in the presence of *aly* suggesting that gene_090 expression is tMAC dependent (HW-C unpublished data). This gene contains several polymorphisms near the transcription start site (TSS) which correlate with its expression.

gene_3L_074: - gene 074 can be found on the 3L chromosome of *Drosophila* and expressed in RAL517 (Zhao et al. 2014) and allele from fully sequenced *aly2* chromosome but not expressed in other lines-RAL304, 307, 357, 360, and 399 (Zhao et al. 2014). Gene_074 show expression in gonads and accessory gland but not in other developmental stages of *Drosophila* **(Figure 1.6)***.* Now, gene_074 is annotated as noncoding RNA called *CR45408* **(Figure 1.6)**. Gene 074 possesses few SNPs compared to 090 gene and there is no correlation between the polymorphism and its expression. Allele of gene_074 expresses from fully sequenced *aly2* chromosome expresses independent of *aly*, suggesting gene_074 expresses independent of tMAC (HW-C unpublished data).

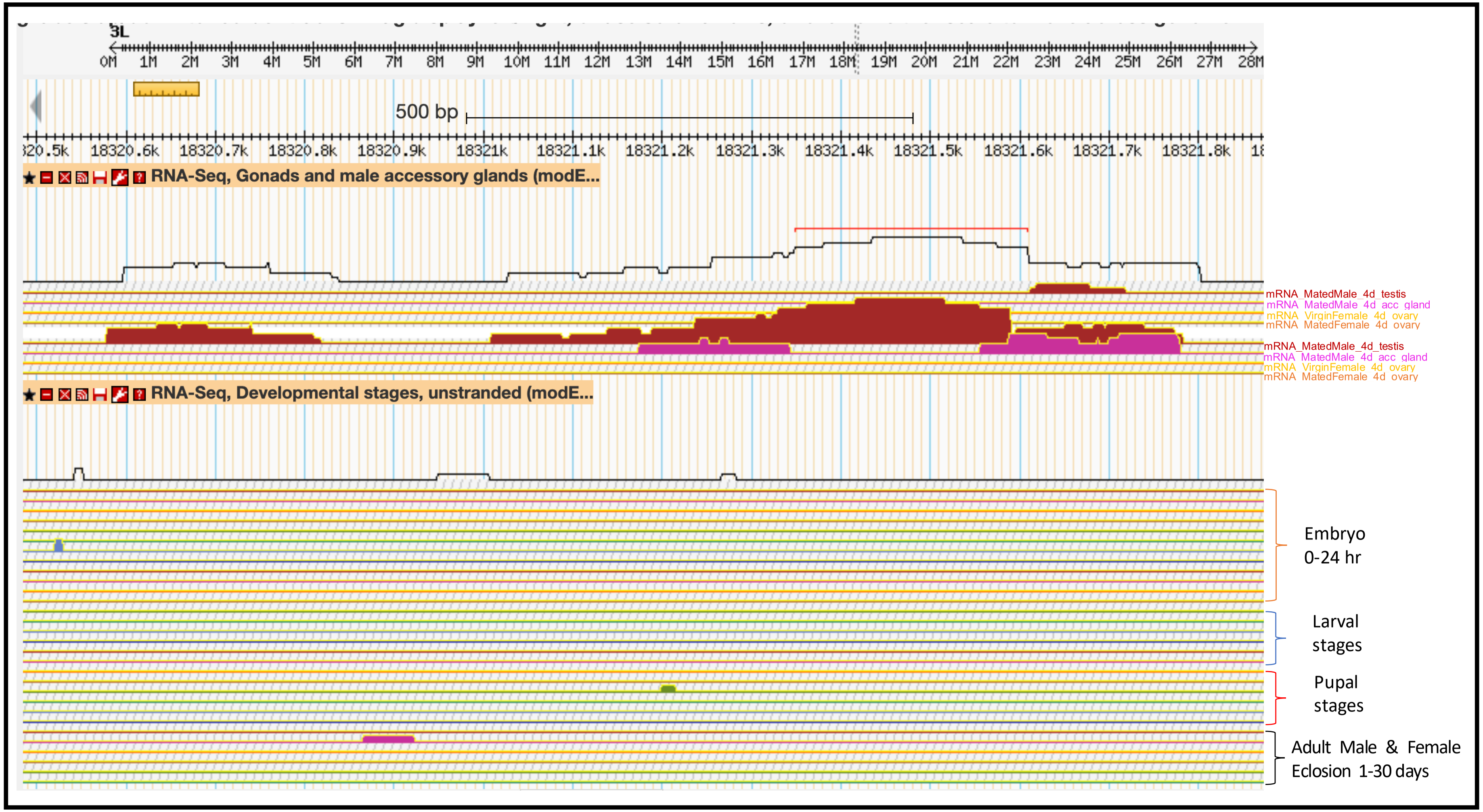

3L
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M 19M 20M 21M 22M 23M 24M 25M 26M 27M 28M
500 bp
320.5k 18320.6k 18320.7k 18320.8k 18320.9k 18321k 18321.1k 18321.2k 18321.3k 18321.4k 18321.5k 18321.6k 18321.7k 18321.8k 18
RNA-Seq, Gonads and male accessory glands (modE...
mRNA_MatedMale_4d_testis
mRNA_MatedMale_4d_acc_gland
mRNA_VirginFemale_4d_ovary
mRNA_MatedFemale_4d_ovary
mRNA_MatedMale_4d_testis
mRNA_MatedMale_4d_acc_gland
mRNA_VirginFemale_4d_ovary
mRNA_MatedFemale_4d_ovary
RNA-Seq, Developmental stages, unstranded (modE...
Embryo 0-24 hr
Larval stages
Pupal stages
Adult Male & Female Eclosion 1-30 days

Figure 1.5:- RNA sequencing data (modENCODE) showing the expression of gene_090 in 4 day old testis (brown) and accessory gland (Pink) (it's contaminated with testis sample) whereas there is no expression in 4 day old virgin females and mated females and other developmental stages of *Drosophila* (Adapted from Flybase).

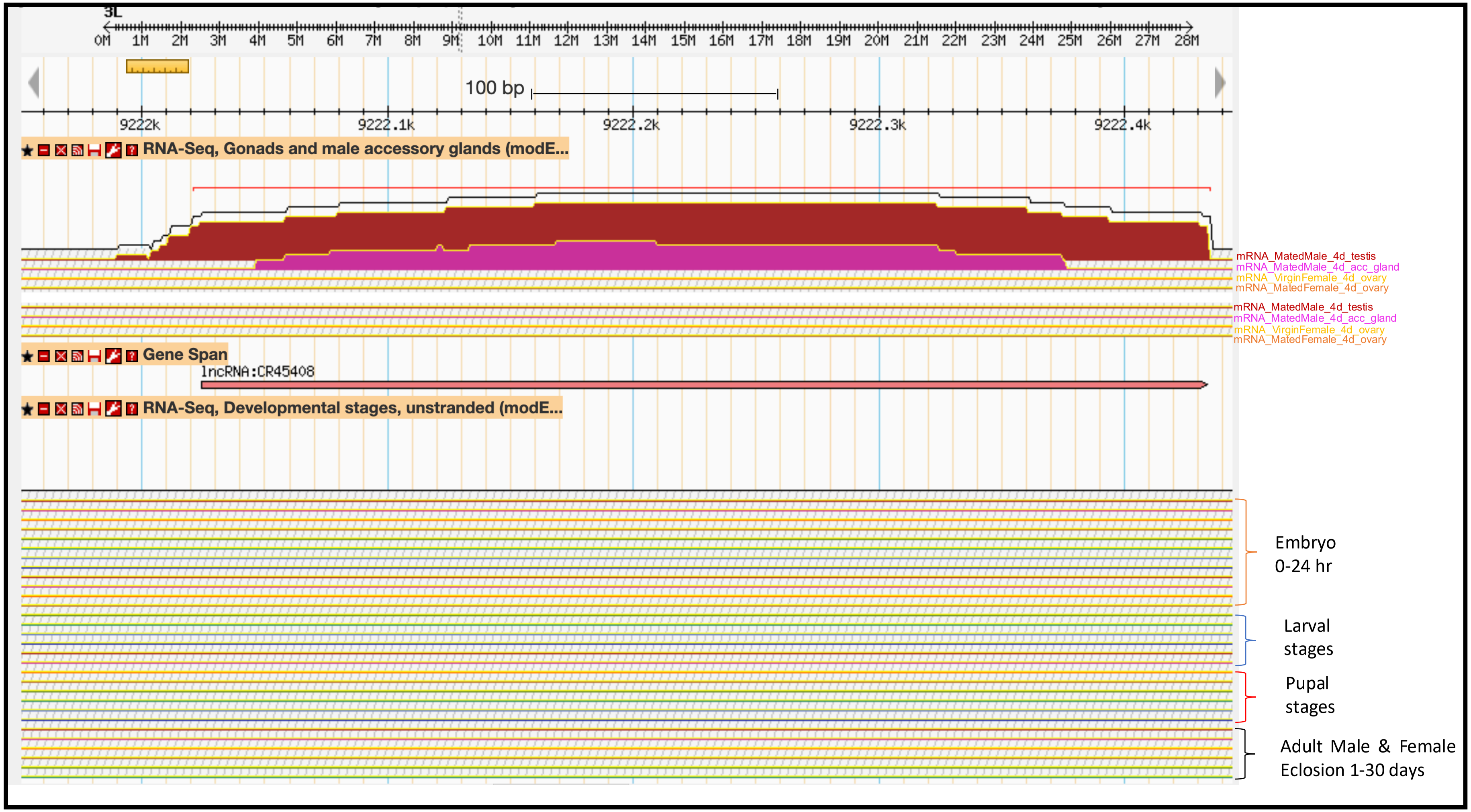

3L
0M 1M 2M 3M 4M 5M 6M 7M 8M 9M 10M 11M 12M 13M 14M 15M 16M 17M 18M 19M 20M 21M 22M 23M 24M 25M 26M 27M 28M
100 bp
9222k
9222.1k
9222.2k
9222.3k
9222.4k
RNA-Seq, Gonads and male accessory glands (modE...
mRNA_MatedMale_4d_testis
mRNA_MatedMale_4d_acc_gland
mRNA_VirginFemale_4d_ovary
mRNA_MatedFemale_4d_ovary
mRNA_MatedMale_4d_testis
mRNA_MatedMale_4d_acc_gland
mRNA_VirginFemale_4d_ovary
mRNA_MatedFemale_4d_ovary
Gene Span
lncRNA:CR45408
RNA-Seq, Developmental stages, unstranded (modE...
Embryo
0-24 hr
Larval
stages
Pupal
stages
Adult Male & Female
Eclosion 1-30 days

Figure 1.6:- RNA sequencing data (modENCODE) showing expression of gene_074 in 4 day old testis (Brown) and accessory gland (Pink) (it's contaminated with testis sample) whereas there is no expression in 4 day old virgin females and mated females and other developmental stages of *Drosophila* (Adapted from Flybase).

1.6. Experimental strategies

In this thesis I wanted to determine the molecular mechanism that drives the expression of two *de novo* genes (gene_074 and gene_090) in *Drosophila* testis and identification and characterization of binding site of C-terminal of Topi protein by SELEX-sequencing. To address these questions, two distinct experimental strategies were used.

1.6.1. Functional analysis of promoter of *de novo* genes

The following figure is an example of expressing and non-expressing alleles showing different promoter regions with respect to the presence of SNP, insertion, and deletion in the promoter regions. Given that, if the allele is expressed in the presence of SNP (T) and insertion or deletion, suggests that presence of SNP (T) is required for the expression of the allele **(Figure 1.7A&C)**. If non-expressing allele gives expression when insertion is replaced with deletion, then deletion is important for expression **(Figure 1.7B&D).** If both C and D versions **(Figure 1.7A&C)** express then additional sequences could be important for expression.

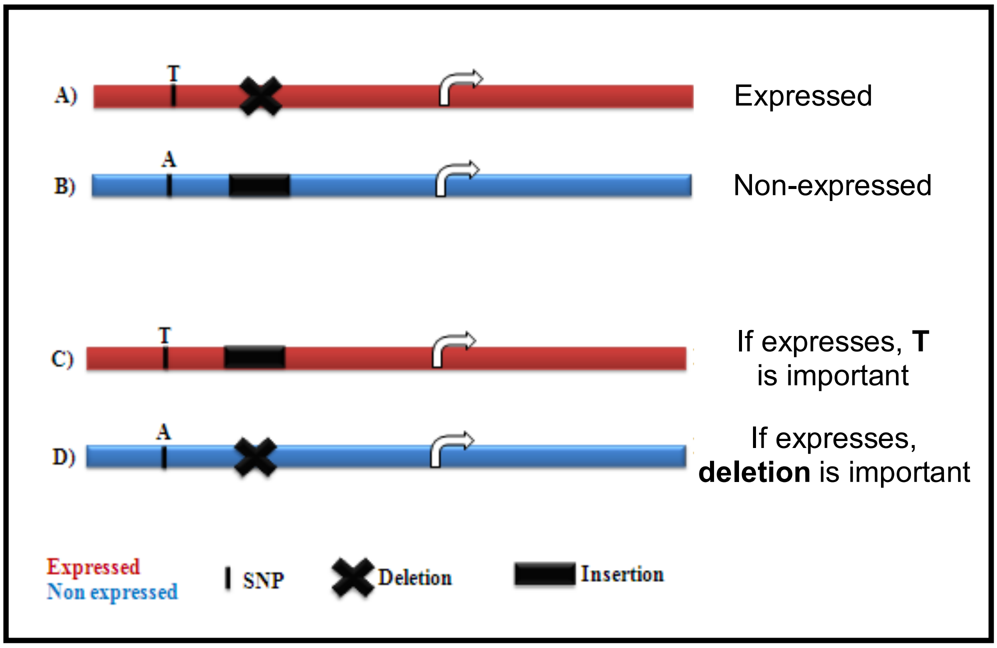


Figure 1.7:- Promoter regions of the expressing and non-expressing alleles.

Studying these two genes would help us identify what region/s of the promoter sequence is/are required for allelic expression. Also, if *de novo* genes are regulated by testis specific complex-tMAC then it would be helpful to identify the differential binding of the tMAC to the promoters of expressing and non-expressing alleles. To analyse the functional aspect of the promoter of these two genes, we would design the individual synthetic DNA fragments for each gene carrying combinations of different SNPs, insertions and deletions from expressing and non-expressing allele. We would clone these designed fragments into multiple cloning site present in vector pCaSpeR-AUG-βGal-attB. Later, these constructs would be injected into attP40 flies which has recombined for site specific recombination and transgenic lines would be generated. This vector lacks the minimal promoter and has *LacZ* gene downstream of the MCS and to analyse these cloned promoter regions we score the beta-galactosidase activity by staining. Its expression will tell us if cloned fragments are capable for driving the expression in testes. Furthermore, to study the mRNA levels of reporter gene, we will use the qRT-PCR and *in situ* hybridization, and X-gal staining. Also, we will study the role of tMAC complex (in the background of different components of this complex) in regulation of gene expression and if these genes are regulated by tMAC, then we will address the differential binding of tMAC to the proximal promoter region of both genes and this will tell us how tMAC complex regulates the gene expression.

1.6.2. *de novo* motif discovery for protein Topi C-terminal by SELEX

Recent advancement in technology enabled the scientists to pursue how and where transcription factors bind in genome and how they regulate the gene expression of target genes. One such cutting edge technology, Systematic evolution of ligands by exponential enrichment (SELEX) allow researchers to determine the binding sites of TFs in the genome of an organism. Common approach of SELEX involve interaction of random DNA oligonucleotides and protein of interest and then purification of protein-DNA complex either by EMSA or immunoprecipitation of protein using suitable antibody against the protein of interest.

Then, elution of DNA molecules from the protein and amplification of the eluted DNA molecules. Rounds of selection of aptamers is user dependent and after all successive rounds of selection, DNA would be sequenced using High-throughput DNA sequencing and discovery of *de novo* motifs **(Figure 1.8)**. To study tMAC-DNA interaction, we will adopt SELEX technique. To produce proteins of interest, subunits of tMAC complex with predicted DNA binding domain will be cloned into expression vector and would be expressed. In addition to above, we will design a 20bp long ssDNA library with a pair of constant regions suitable for use as PCR primer target sites flanking the variable insert and convert into dsDNA library using Klenow fragment. For each round of SELEX, we will prepare the protein-Ni-NTA agarose beads complex and allow it to bind to newly synthesised dsDNA library. Bound DNA would be eluted and amplified by Emulsion PCR. Emulsion PCR would allow us to amplify individual DNA molecules with fewer biases than found in conventional PCR **(Figure 1.8)**.

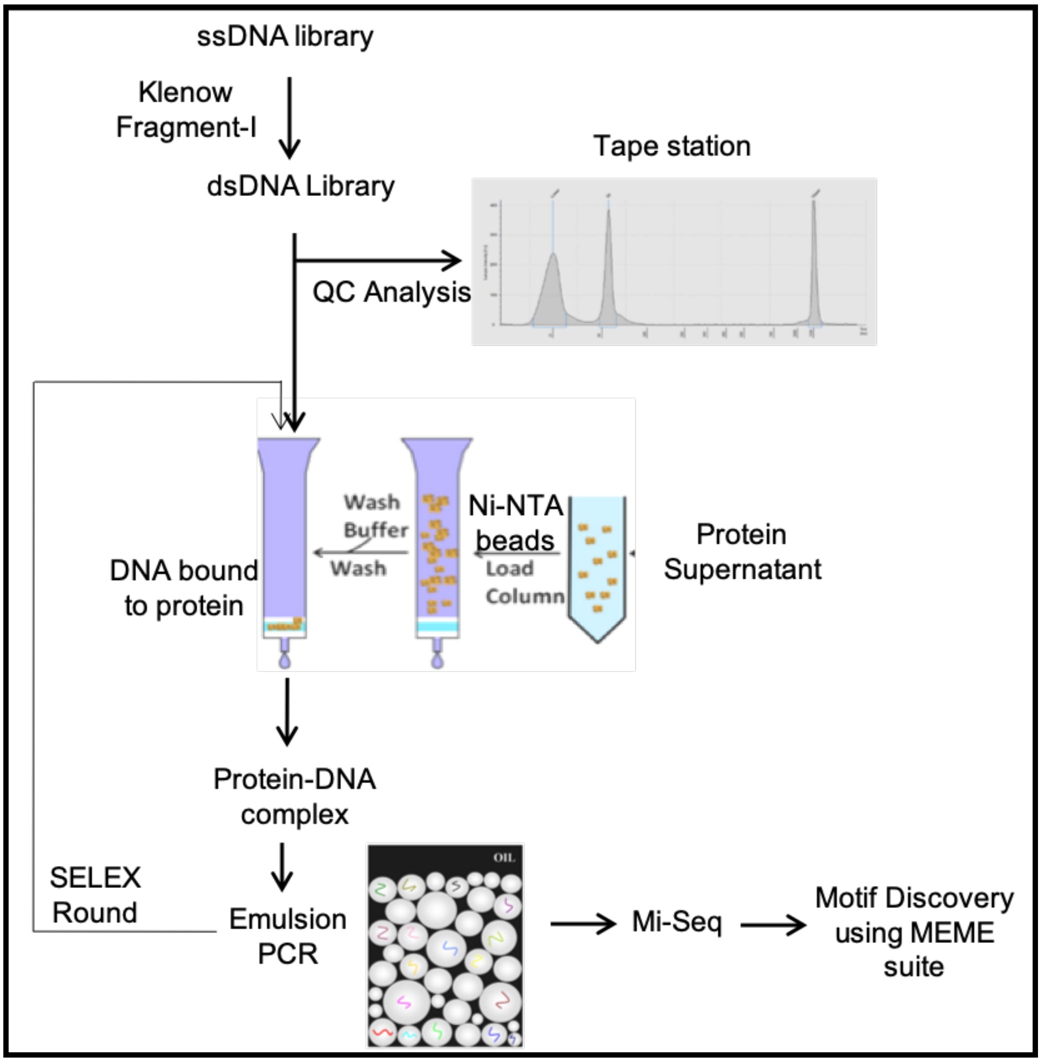


Figure 1.8:- Our approach for SELEX-sequencing. Design and synthesis of dsDNA library and quality check by Tape-station. Incubation of synthesised dsDNA library with protein bound to Ni-NTA-agarose beads. Elution of Protein-DNA complex and amplification of eluted DNA by emulsion PCR, sequencing by Mi-sequencing and motif discovery.

1.7. Aims of this thesis

In the following chapters I will present:

1. The molecular mechanisms underlying evolution of testis-specific expression of *de novo* genes in *Drosophila*
    - designing of synthetic DNA fragments of gene_074 and gene_090
    - generation of transgenic lines for *de novo* genes
    - functional analysis of promoter regions of gene_074 and gene_090
    - expression analysis of reporter gene in testes from transgenic lines
    - addressing the mechanism of gene expression regulation
2. Over-expression and characterization of tMAC subunits
    - cloning of tMAC subunits
    - over-expression of tMAC subunits
    - characterization and purification of expressed proteins
3. tMAC-DNA binding analysis by SELEX
    - design of ssDNA library and synthesis of dsDNA library
    - characterization of synthesized library
    - identification of DNA binding site for Topi C-term
    - characterization of DNA binding site by EMSA

Chapter 2. Materials & Methods

2.1. Molecular biology methods

2.1.1 Genomic DNA extraction

Individual flies for PCR were smashed well in 50µl squishing buffer (10mM Tris pH8.2, 1mM EDTA, 25mM NaCl, 0.2µg/ml Proteinase K make up to 100ml with H_2O) and incubated at room temperature for 30 minutes. The proteinase reaction was then inactivated by incubating samples at 95^0C for 3 minutes. Samples were centrifuged at 13,000rpm for 5 minutes and clear solution of genomic DNA was transferred to new Eppendorf tube and stored at -20^0C until needed.

2.1.2 Polymerase Chain Reaction (PCR)

All PCR reactions were performed using *Thermus aquaticus* (Taq) DNA polymerase and carried out in a Biometra T3 thermocycler. Each PCR reaction consisted of a standard mix:

1.0µl gDNA template (1ng)

0.4µl dNTPs (10mM each)

0.5µl 5'primer (10pM)

0.5µl 3'primer (10pM)

2.0µl 10X buffer

0.2µl Taq DNA polymerase

15.4µl H_2O

20.0µl total volume

2.1.3 RNA extraction and cDNA synthesis

Wild type testes were dissected in 1x testis buffer (183mM KCL, 47mM NaCl, 10mM Tris pH 6.9) and immediately moved to lysis buffer provided by RNAqueous®-Micro Kit (Life Technologies) and stored at -80 until needed. Total RNA extraction was performed as per manufacturer's instructions until elution step and RNA was eluted in 2x6.5µl of elution buffer. cDNA was synthesised by the Superscript III first strand synthesis system (Life Technologies). For 20µl reaction, total extracted RNA from a pair of testis was incubated with 1µl of oligodT (50µM), 1µl of dNTPs (10mM) and reaction mix was incubated at 65^0C for 5min. After incubation, the reaction mix was cooled down by incubating on ice for at least 2 minutes. For first strand synthesis, the reaction mix was incubated with 4µL of 5X First strand buffer (50mM Tris-HCL-pH 8.3, 75mM KCL, and 3mM MgCl2), 1µl of 40

Units of RNaseOut^{TM}, 1µl of 0.1M DTT, and 1µl of 200Units of Superscript III RT (Life technologies) at 50^{0}C for 1 hour. The reaction was stopped by incubating at 70^{0}C for 15 min, and further reaction was diluted to 60µl by adding 40µl of dH$_2$O and stored at -20^{0}C. Gene specific primers were used to amplify the cDNA (see Appendix 1 **)** and correct size of each amplicons were confirmed by 0.8-1.5% (according to the products) agarose gel electrophoresis. Amplified products were purified by using Gel extraction kit from Qiagen (Cat No. /ID: 28704) as per manufacturer's instruction.

2.1.4. Designing and molecular cloning of synthetic DNA fragments of *de novo* genes

2.1.4.1 Designing of synthetic DNA fragments for gene_074 and gene_090

In order to analyze the promoter sequences of both the genes, we designed the different DNA fragment carrying a combination of SNPs, insertion, and deletion from both expressing and non-expressing alleles of both genes.

For gene_074:- Total three fragments were designed:-

074_ref- It has a reference genomic sequence of *Drosophila*.

074_aly- It consists the sequence of allele of gene_074 from the aly2 chromosome (stock HWC1) with SNP upstream of the TSS.

074_517- It has the sequence from the highly inbred line of *Drosophila* RAL_517 with SNP at upstream and downstream of the TSS.

For gene_090:- For gene_090, total 17 different fragments were designed. All constructs has the sequences from -302bp to +133bp relative to the reference genome:-

090_A- It has the sequence from -302bp to +133bp from one of the expressed alleles.

090_B- It has the sequence from -302bp to +133bp from one of the non-expressed alleles.

090_C- It has the sequence from -302bp to +133bp with the SNPs derived from the expressed alleles but both insertions found in the non-expressed alleles.

090_D-It has the sequence from -302bp to +133bp with the SNPs derived from the non-expressed alleles but both deletions found in the expressed alleles.

090_E- It has the sequence from -302bp to +133bp with the SNPs, and the first deletion derived from the expressed alleles but the second insertion found in the non-expressed alleles.

090_F- It has the sequence from -302 to +133 with the SNPs, and the second deletion derived from the expressed alleles but the first insertion found in the non-expressed alleles.

090_G- It has the sequence from -302 to +133 with both deletions from the expressed alleles. All the SNPs are from the expressed alleles except for the single SNP that is in the potential tce. ie has a single base pair change compared to the expressed sequence version.

090_H- It has the sequence from non-expressed allele first deletion is from expressed allele and SNPs and second insertion are from non-expressed allele.

090_I- It has the Sequence from expressed allele, with insertion at indel site 1 of CTACATA (sequence from upstream)

090_J- It has the sequence from non-expressed allele, with insertion at indel site 1 of CTACATA (sequence from upstream)

090_K- Sequence from expressed allele, with insertion at indel site 1 of CTTTTGC (sequence from downstream)

090_L- Sequence from non-expressed allele, with insertion at indel site 1 of CTTTTGC (sequence from downstream)

090_M- Sequence from expressed allele, with insertion at indel site 1 of TAGAGAT (randomization of the current insertion sequence)

090_N- Sequence from expressed allele, with insertion at indel site 1 of CCATTAT (inversion of the current insertion sequence)

090_O- Sequence from expressed allele, with deletion at indel site 1 but insertion of ATAATGG (insertion sequence for indel1), inserted 20 bp upstream of current site (2 DNA helix turns).

090_P- Sequence from expressed allele, with deletion at indel site 1 but insertion of ATAATGG (insertion sequence for indel1), inserted 15 bp upstream of current site (1.5 DNA helix turns).

090_Q- It has the sequence from the expressed allele with two bases are altered from “AC” to “CA” at indel-I site.

2.1.4.2. Insertion of the attB site into pCaSpeR-AUG-bGal

attB and attP site specific recombination system is been extensively used to incorporate DNA fragment of interest in the genome and φc31 recombinase enzyme is responsible for this recombination. To achieve this, DNA fragment of interest should possess attB site and recipient genome attP site. In our case, we adopted attB site and attP40 site specific recombination system. For retrieval and generation of attB site, 1µg of pUAST vector was treated in the presence of 1X reaction Buffer H (#R008A, 10mM Tris-HCl (pH 7.3), 300mM NaCl, 0.1mM EDTA, 1mM DTT, 0.15% Triton X-100, 0.5mg/ml BSA, 50% glycerol) and 1ul (10U) of *Pst* 1 restriction enzyme (R6111) for 2hr at 37^{0}C and reaction was stopped by incubating the reaction mixture at 80^{0}C for 20 minutes. Further, the digested fragment was subjected to purification using agarose gel electrophoresis and Qiagen Gel Extraction Kit (Cat No. /ID: 28704) and amplified using primer flanking *Pst* 1 site (Primer-attB+*Pst*1) (See Appendix 1). The amplified product was purified by gel extraction kit as above mentioned and sequenced.

2.1.4.3. Molecular cloning using Gibson assembly

Gibson assembly is a method of molecular cloning which allow us to join multiple DNA fragments using single reaction (Gibson et al. 2009). This assembly is made up of few components specifically exonuclease, DNA polymerase, and DNA ligase. For cloning of the attB site, in 20µl reaction, 1µg of pCaSpeR-AUG-βGal vector was digested by 1µl (10U) of *Pst* 1 restriction enzyme using buffer above mentioned for 2hrs at 37^{0}C. Cloning of the attB site was done by using Gibson assembly. For this purpose 1:4 ratio of vector DNA to PCR fragment(attB site) was used and 20µl of the reaction mixture with 50ng of purified vector, attB site PCR fragment, 2X of Gibson Assembly Master Mix, and volume adjusted by deionized water incubated at 50^{0}C for 15 minutes. 5µl of the reaction mixture was used for transformation into 50µl of DH5α cells (NEBC2987H), tapped gently and kept on ice for 2 minutes. Transformed cells were spread onto Ampicillin containing Agar Petri dish and allowed to grow for 14-16hrs at 37^{0}C. The Colony PCR was performed by carefully selecting a single colony and PCR was carried out by using primer flanking the insert to confirm the insert. The amplified product was sent off for the sequencing and respective colony was used for the maxi preparation of plasmid.

2.1.4.4. Cloning of de novo genes DNA fragment

The individual fragment was diluted with deionized water and used for the cloning purpose. Total 50 ng of pCaSpeR-AUG-βGal-attB plasmid was digested with *EcoR*I and *BamH*I restriction enzymes to excise the *dj*lΔ52 fragment from the plasmid (the procedure is given above). Cloning of each fragment was performed using Gibson assembly and the procedure is given above. Each insert was confirmed by colony PCR with a primer pair flanking the multiple cloning site. Positive clones were verified by sequencing.

2.1.5. Molecular cloning of tMAC components

2µg of pET28a (+) vector was digested by 2µl of NdeI (20,000U/ml) and HindIII (20,000U/ml) each in 1X Cut Smart buffer (10X stock, from NEB) and final volume of 20µl was made with dH_2O. The reaction was mixed with gentle pipetting and incubated at 37^0C for 2 hours. Reaction was stopped by heat inactivation at 70^0C for 20 min and digested vector was resolved onto 0.8% agarose gel. The digested vector band was excised form the gel and purified by Gel extraction kit from Qiagen (Cat No. /ID: 28704). For dephosphorylation of 5' ends of vector, 20µl reaction was set up with vector, 1X Antarctic Phosphatase reaction buffer (stock 10X), 1µl (5units) of Antarctic Phosphatase and final volume was made by dH_2O. The reaction was mixed by gentle pipetting and incubated at 37^0C for 30min. The reaction was stopped by heat-inactivation at 80^0C for 2 min and cooled down on ice for ligation. Cloning of Full length Comr, Topi, Tomb, Achi/vis, Kruppel as well as DBD of all the proteins (except Kruppel) was performed by T4 DNA ligase or Gibson assembly in pET21a vector. Cloning of Topi C-term and Achi/vis FL was performed using Gibson assembly. For Gibson assembly cloning, 20µl reaction was set up with 1:4 vector (50ng) to inset molar ratio, 10µl of Gibson Assembly Master Mix (2X) and final volume was made up with dH_2O. The reaction was mixed with gentle pipetting and incubated at 50^0C for 15 min. 5µl of ligation mix was transformed into 50µl DH5alpha bacteria, mixed by tapping the tube, placed it on dry water bath at 42^0C for 30 sec, and immediately placed on ice for 2 min. Later, 300µl of SOC outgrowth medium (previously incubated at 37^0C) was added to the bacterial tube and incubated at 37^0C for 1hr with vigorous shaking at 250rpm. Incubated bacteria was plated onto Ampicillin containing agar plate and incubated at 37^0C for overnight. T4 DNA ligation was performed for the rest of the fragments. For each insert, 20µl reaction was set up with 1:3 molar ratios (50ng vector) of digested vector to insert, 1X T4 DNA ligase buffer, 1µl of T4 DNA ligase and final volume was made up with dH_2O. The reaction mixed by gentle pipetting and incubated at room temperature for 2 hour. The reaction was stopped by heat-inactivation at 65^0C for 10 minutes. 5µl ligation mix was transformed into DH5alpha bacteria, incubated at 37^0C for overnight. On the next day, colony PCR was carried out using T7 (promoter and terminator) primer set to check insertion of desired fragments and sequences were confirmed by Sanger sequencing.

2.1.6. ssDNA library design

A Hand-mixed ssDNA library of 20nt with primers flanking both ends was ordered from Integrative DNA technology (IDT). Primer sequences Selex_F 5'-GGTATTG AGGGTCGCATC-3'-N20-Selex_R 5'-AGAGGAGAGTTAGAGCCATC-3' was adopted from Murphy, M.B., et al*,* 2003. The library was reconstituted in RNase/DNase free water to 100µM concentration.

2.1.7. dsDNA library synthesis and purification

For a 50µl reaction, 5µl of 100µM ssDNA template, 7.5µl of 100µM reverse primer, 4µl of 2.5mM dNTPs, 5µl of 10X Klenow buffer, and 26.5µl of sterile ddH_2O was denatured in boiling water for 5 minutes and slowly cooled down at room temperature. 2µl of Klenow enzymes (5U/µl, New England Biolabs) was added to the reaction and incubated at 37^0C for 20 min. Klenow enzyme was inactivated by incubating the reaction mix at 75^0C for 10 minutes. Newly synthesised dsDNA library loaded to Microcon YM-30 (Millipore) and spun at 5000g for 15 minutes and to recover the concentrate column was inverted and spun at 1000g for 2 minutes. Filtered and concentrated dsDNA library was detected by 2% agarose gel and quantify by Qubit (Invitrogen).

2.1.8. Systematic evolution of ligands by exponential enrichment (SELEX)

For preparation of protein-beads complex (see section 2.3), 0.2ml Ni-NTA beads with 50% slurry was added to the 1ml topi C-term supernatant and rotated for 1hr at 4^0C. Beads were briefly spun down at 1000rpm and supernatant was removed. Then beads were washed 3X10min with 1ml of binding buffer (50mM Na_2HPO_4-NaH_2PO_4, 150mM NaCl pH7.8) by gentle mixing. Beads were pelleted down by short spin and supernatant was stored from each wash. Later, the beads were rinsed with wash buffer (20mM sodium phosphate, 500mM NaCl pH 6.0) and washed once for 10 minutes and beads bound protein complex was stored at 4^0C. 1% of beads was heated to 100^0C and ran onto SDS-PAGE using 1X SDS running buffer. Another 1% beads were used to check the protein concentration. Protein bound beads were eluted by 20 µl of imidazole elution buffer (20mM sodium phosphate, 500mM NaCl, and 500mM imidazole pH 6.0) and the protein concentration was checked by Bradford reagent at 595nm. For the first round of SELEX, 100pM beads bound protein was incubated with 1nM dsDNA library in 1ml of binding buffer (20mM sodium phosphate, 500mM NaCl pH-7.8) and incubated at room temperature with rotation for 30 minutes. Later, beads were washed 3 times with 0.1% PBS-T by inverting and the beads were then collected by brief spin at 1000rpm. The protein-DNA complex was eluted from Ni-NTA beads with 20µl of SELEX elution buffer (20mM Tris pH 7.5, 500mM imidazole) and transferred into PCR tube. Eluted DNA was amplified by Emulsion PCR and the amplified product was used for next round of SELEX. For rounds two and three of SELEX, 75pM of beads bound to protein was used and incubated with 1nM of DNA in 1ml of binding buffer.

2.1.9. Emulsion PCR

2.1.9.1. Water-in-oil preparation

Water-in-oil emulsion for PCR was prepared according to (Shao et al. 2011). The oil phase, prepared freshly each day, was composed of 4.5% Span 80, 0.4% Tween 80, and 0.05% Triton X-100 in mineral oil. The aqueous phase of water-in-oil emulsion was PCR mixture. Water-in-oil emulsion was prepared by adding 100µl of ice-cooled PCR reaction mixture gradually (in 10 aliquots of 10µl over 2 minutes) to 200µl of oil phase in 2ml round-bottom eppendorf tube whilst the mixture was continuously stirred at 1500 rpm for 5min before PCR cycling.

2.1.9.2. PCR

100µl of conventional PCR mixture was prepared as below: 3.2ul of each primer (10µM stock), 50µl of One Taq 2X Master Mix, 1ng of dsDNA template, and final volume was adjusted by dH_2O. PCR amplification was performed on Chromo4 instrument (MJR). PCR procedures were carried out under the following cycling conditions: 94^0C for 2min, 35 cycles of 94^0C for 30sec, 56^0C for 30sec, and 72^0C for 30sec.

2.1.9.3. Recovery of the reaction mixture

The water-in-oil emulsions from the PCR tube were pooled in 2ml tube and spun at 9000g for 5 min. The oil phase was removed and two volumes of water-saturated ether were added to the intact concentrate left at the bottom of the vial. The mixture was then vortexed, centrifuged briefly, and ether phase was removed. The aqueous phase was washed two times with ether and dried at room temperature for 30 min. The recovered product was quantified using a Qubit (Invitrogen)

2.1.10. Electrophoretic mobility shift assay (EMSA)

Single stranded positive and negative unlabelled probes (19bp each) were designed based on the results from Mi-sequencing and *de novo* motif discovery using MEME suite. An equimolar concentrations of each ssDNA probe was mixed with its complementary ssDNA, incubated at 95^{0}C for 5 minutes and slowly cooled down at room temperature for 60 minutes. EMSA reactions were set up as follow (25µl): 4.5µl of %X DNA binding buffer (50mM Tris-HCl, pH7.5, 125mM NaCl, 2.5mM EDTA, 15mM $MgCl_{2,}$ 5% glycerol, 2.5mg/ml BSA, 0.25% NP-40, 0.05mM DTT), 10µl of purified Topi C-terminal protein (100pM), 4µl dsDNA unlabelled probes (positive and negative, 1nM), and the final volume was made up by dH_2O. The reactions were incubated at room temperature for 60 min. To assess the binding, reactions were fractionated by 1.5% agarose in 0.5X TBE (10X TBE: 1M Tris base, Boric acid, and 0.02M EDTA). The gel was stained with ethidium bromide (100µg/ml) for 15 minutes at room temperature and visualised with UV-transilluminator.

2.2. Insect methods

2.2.1. Maintenance of Drosophila stocks

All *D. melanogaster* stocks were maintained at 25^{0}C on a standard artificial media made up of cornmeal, sucrose, yeast, and agar medium. Standard genetic techniques were used in manipulating the *Drosophila* genotypes. Stocks of flies used were- *w1118, w+; if/cyo, yw;nos-φ31;attp40, aly2 rede/TM6C, aly5 red/TM6C.*

2.2.2. Preparation of fruit juice agar plates

3g agar was added to 100ml of H_2O and boiled while 3g sucrose was dissolved in 30ml cranberry juice and boiled. When both had cooled they were mixed and 2ml nipagin (10% nipagin in 95% ethanol) added. This solution was then poured into small plastic cups and when set stored at 4^{0}C.

2.2.3. Collection of *Drosophila* embryos

Flies of the desired genotype were placed in empty plastic bottle with fruit juice agar plate sealed to the opening of the bottle (with smear of yeast paste). Flies were kept at 25^{0}C. After each 45 minutes embryos were collected and used for microinjection.

2.2.4. Germline transformation of *Drosophila* embryos

To successfully integrate a transgene into *Drosophila* germline, a construct containing the transgene must be injected into blastoderm embryos (aged from 45-90 minutes). For dechorionation, double sided sticky tape was placed on a glass slide and thin copper wire was placed onto other side of the tape. *yw;nos-φ31;attp40* embryos were collected and placed on the other side of tape with copper wire. Another slide was with double sided tape placed onto embryo containing slide and pressed against each other and separated, exposing dechorionated embryos. Embryos were then transferred onto a glass coverslip coated in 3M magic tape glue dissolved in heptane and aligned with anterior-posterior axis in parallel and orientated in the same direction. The embryos were briefly desiccated for 5-15 minutes and covered with 10s Voltalef oil (Sigma).

Femtotip II needles were loaded with 2µl of the relevant construct (~100ng/construct) and embryos were injected into the posterior end using an Eppendorf Transjector 5246 and micromanipulator connected to a Nikon microscope.

Injected embryos were left to recover under 10s Voltalef oil on fruit juice agar plates with yeast at 18^{0}C. Surviving larvae were recovered and cultured at 25^{0}C on standard diet.

2.2.5. X-gal staining

Testes were dissected in testis buffer and transferred to 24 well culture dish. Testes were fixed using 1% glutaraldehyde (made in 1xPBS) for 15 min and rinsed with 1X X-gal buffer (Stock solution 5X X-gal buffer- 90ml of 0.2M $Na_2HPO_{4,}$ 35ml of 0.2M $NaH_2PO_{4,}$ 2.5ml of 1M $MgCl_{2,}$ 75ml of 5M NaCl and make up the volume of 500ml with dH_2O) and kept in this buffer for at least 30 min. Staining solution was prepared using 1X X-gal buffer, 50mM $K_4Fe(CN)_6$, 50mM $K_3Fe(CN)_6$, and 20% of X-gal (in DMF), vortexed and kept at 37^{0}C. Later, testes were incubated with this staining solution at 37^{0}C overnight (sometimes need to keep longer till colour develops). Stained testes were washed in 1X X-gal buffer and mounted in 85% glycerol.

2.3. Protein methods

2.3.1. Protein expression and validation by SDS-PAGE

For protein over expression, each construct was transformed into BL-21dm (*E.coli*) expression bacteria and allowed to grow on to LB *Kanamycin* agar plates overnight at 37^{0}C. A single colony was picked and incubated in 5ml of LB broth with 50µg/ml of

Kanamycin for overnight at 37^0C. Next day, 1ml of culture from 5ml of culture was incubated in 100ml of LB broth with 50µg/ml Kanamycin till O.D reached to 0.4-0.6 at 590nm. For protein induction, these cultures were incubated with 0.5mM or 1mM IPTG for 3hrs at 25^0C with shaking at 230 rpm. Cells were collected by centrifuging at 5000rpm for 10 min at 4^0C. The cell pellet was suspended in 4ml of binding buffer (20mM sodium phosphate, 500mM NaCl pH-7.8) with 1mg/ml lysozyme and 1X Protease inhibitor cocktail (Roche) and incubated on ice for 30 min. Later, the mixture was incubated on shaker for 10 minutes at 4^0C. Mixture was incubated with 1% of Triton-X 100, 5µg/ml of DNase and RNase for another 10 minutes on shaker at 4^0C. Lysed cells were spun at 15,000rpm for 30 minutes at 4^0C. The supernatant was collected in new vials and both pellet and supernatant were stored at -80^0C. Protein over expression was checked on 10-12% Acrylamide-bis Acrylamide gel in 1X SDS running buffer. For Topi FL, cultures were incubated for 4hrs at the same conditioned as mentioned above and protein over expression was checked on 10% Acrylamide-bis Acrylamide gel.

2.3.2. Purification of His-tagged protein

Recovered supernatant from earlier stage was mixed with 1ml of Ni-NTA agarose beads (Qiagen) and rotated for 1hour at 4^0C to allow the His-tagged protein to bind to the beads. Later, samples were loaded onto the column and washed 3 times with wash buffer (20mM sodium phosphate, 500mM NaCl pH-6.3). Protein was eluted from column in 4x500µl fractions in elution buffer (20mM sodium phosphate, 500mM NaCl pH-5). Samples were analysed for quantity and purity by SDS-PAGE.

2.4. Cell biology methods

2.4.1. Synthesis of DIG-Labelled RNA antisense probe

100-200 ng PCR product of the *LacZ* gene as used as template in a 20µl reaction with 2µl 10X DIG RNA labelling mix (Roche), 2µl 10X transcription buffer and 2µl RNA polymerase T7 (20U/µl) (Roche). The reactions were incubated at 37^0C for 2 hours and the probes were then hydrolysed in 80µl DEPC treated 200mM carbonate buffer (40mM $NaHCO_3$, 60mM Na_2CO_3) for 30 minutes per 1kb of length. This was neutralised by the addition of 100µl of DEPC treated 200mM sodium acetate and the probes precipitated with 2.5 volumes of ethanol at -20^0C for about 1 hour. The samples were then centrifuged at 4^0C for 30 minutes at 13000rpm, the pellet washed with 70% ethanol and air dried. Probes were resuspended in 200µl DEPC treated H_2O and stored at $–80^0$C.

2.4.2. Dissection and fixation of testes

Testes were dissected from young males (0-1 days old) in testis buffer (183mM KCl, 47mM NaCl, 10mM Tris pH6.9) using a dissecting microscope and fixed for 30 minutes in 4% paraformaldehyde in HEPES buffer (100mM HEPES pH6.9, 2mM $MgSO_4$, 1mM EGTA). Then washed 3 X 5 minutes in PBST (PBS + 0.1% Tween 20). Testes were treated with proteinase K (50µg/ml) for 5 minutes and the digestion was stopped with 2mg/ml glycine for 2 minutes, followed by 2X5 minutes washes in PBST. Testes were re-fixed in 4% paraformaldehyde in HEPES for 20 minutes. They were then washed 3X10 minutes in PBST and 1X10 minutes in 1:1 PBST: hybridisation buffer (HB: 50% formamide, 5X SSC, 100µg/ml sonicated salmon sperm DNA, 50µg/ml heparin, 0.1% Tween 20 and adjusted to pH 4.5 with citric acid).

2.4.3. RNA probe hybridisation

Testes were pre-hybridised for at least 1 hour at 65^0C in hybridisation buffer. RNA probes were denatured by heating at 80^0C for 10 minutes, cooled on ice for a few minutes and added to the samples for overnight hybridisation at 65^0C. The following hybridisation washes were carried out in a water bath as listed in Table 2.1.

Wash buffer	Duration of each wash	Number of washes	Temperature
Hybridisation buffer	20-30 minutes	At least 6 changes	65^0C
4:1 HB:PBST	15 minutes	Once	Room temperature
3:2 HB:PBST	15 minutes	Once	Room temperature
2:3 HB:PBST	15 minutes	Once	Room temperature
1:4 HB:PBST	15 minutes	Once	Room temperature
PBST	15 minutes	Twice	Room temperature

Table 4:-Post hybridisation washes for RNA *in-situ*

Testes were then incubated overnight at 4^0C in a 1:2000 dilution of pre-absorbed alkaline phosphatase conjugated anti-digoxigenin antibody in PBST (antibody preabsorbed against fixed ovaries). Samples were washed 4 X 10 minutes in PBST and then 3 X 5 minutes in a high pH buffer (HP: 100mM NaCl, 50mM $MgCl_2$, 100mM Tris pH9.5, 0.1% Tween 20). The colouration reaction added to the testes consisted of 4.5µl of Nitro blue tetrazolium chloride (NBT) (Roche) and 3.5µl of X- phosphate/5-Bromo-4-chloro-3-indolyl-phosphate (BCIP) (Roche) in 992µl of HP buffer (to give a final concentration of

0.175 mg/ml of BCIP and 0.45mg/ml of NBT per 1ml). The reaction was left to develop in the dark, and on average took between 20 minutes to 2 hours to develop. The reactions were terminated by 3 X 5 minutes washes in PBST.

2.4.4. Mounting of testes

The samples were dehydrated through a series of ethanol washes (30% ethanol in water, 50%, 70%, 90% and 100% ethanol). Each wash was carried out twice. After the dehydration steps, samples were transferred to glass blocks and incubated in 1:1 ethanol:methylsalicylate for 15 minutes followed by incubation in 100% methylsalicylate. The methylsalicylate was replaced with Gary's Magic Medium (GMM) (1.6g/ml Canada balsam in methylsalicylate). Testes were then mounted on glass slides in GMM. Slides were analysed using Normaski Differential Interference Contrast (DIC) microscopy using an Olympus BX50 microscope.

2.5. Bioinformatics techniques

Multiple sequence alignments were carried out by online tool Muscle MSA (http:///www.ebi.ac.uk/Tools/msa/). The phylogenetic trees were drawn using MEGA software with Maximum likelihood statistical test and Bootstrap phylogeny test with 1000 replications. For SELEX-sequencing, sequencing reads were processed using online tool Galaxy (https://usegalaxy.org/). MEME suite v5.4.1(http://meme-suite.org/) was used for *de novo* motif search using default parameter with discriminative search mode and motif comparisons were carried out using online tool TOMTOM with default parameter (http://meme-suite.org/tools/tomtom). FIMO from MEME suite was used to find the occurrence of the motifs with default parameter and *p*-value was adjusted (http://meme-suite.org/tools/fimo).

Chapter 3. Results

Molecular and functional analysis of the promoter of gene_090

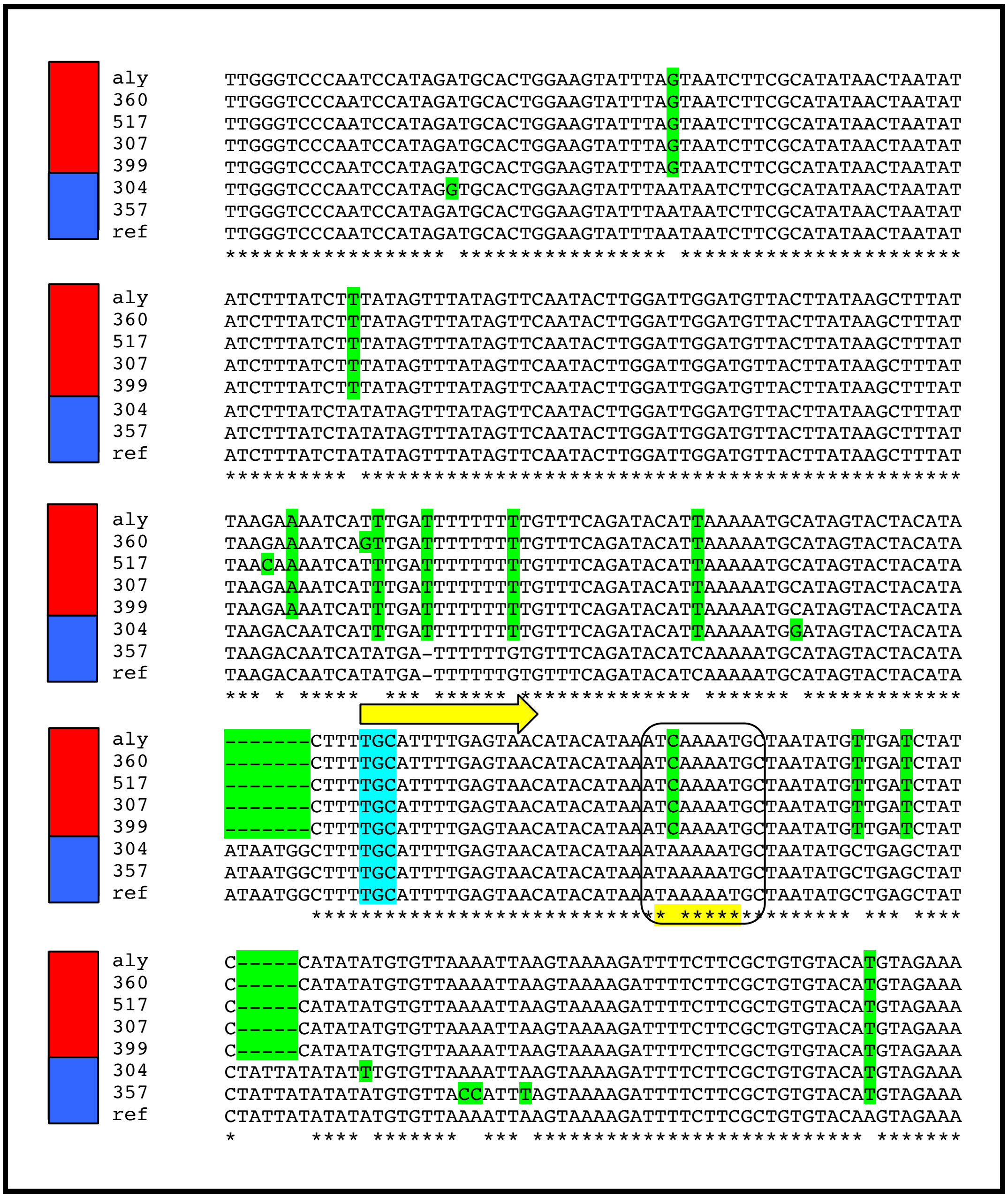
3.1. Abstract

Understanding the origin of new genes in the genome is primary goal of evolutionary biology. Studying the recently evolved *de novo* genes and how they gain expression can be informative at elucidating the molecular mechanisms that give rise to gene expression. This chapter depicts the molecular and functional analysis of the promoter of gene_090 in *D. melanogaster* (Zhao et al. 2014). A gene_090 is interesting to study because promoter region of expressed and non-expressed allele of this gene is contrasted by several single nucleotide polymorphisms (SNPs) and indels. A series of constructs revealed that 7bp indel is necessary and sufficient to convert natural non-expressing allele into an expressing allele.

3.2. The molecular properties and evolution of gene_090

The sequence alignment of expressed and non-expressed allele of gene_090 show 12 SNPs upstream as well as downstream of the transcription start site **(Figure 3.2.1)**. This gene also has two indels, the first 7 bp just upstream of TSS and the second 5 bp downstream of TSS **(Figure 3.2.1)**. Zhao et al. established that transcripts with >2 FPKM were considered as *de novo* transcripts (explained in section 1.5.4) and gene_090 was found to be expressed in RAL- 307 (3.97 FPKM), 360 (3.34 FPKM), 399 (5.92 FPKM), and 517 (12.74 FPKM) lines but was not expressed in RAL-304 (0.03 FPKM), 357 (0.00 FPKM) lines. An allele of gene_090 from *aly* chromosome was tested in our lab and was found to be expressed (HW-C unpublished data).

A study identified a core promoter sequences of 190 genes specifically expressed in testis that have a 10 bp A/T-rich motif that is identical to the translational control element (TCE). Mutation of TCE reduced the transcription of reporter gene indicating that the TCE is important but not essential for activation of transcription (Katzenberger et al. 2012). Intriguingly, gene_090 has a predicted translational control element (TCE) (Katzenberger et al. 2012) which possibly also regulates the transcription of the gene **(Figure 3.2.1)**. A phylogenetic tree was created using promoter regions of gene_090 and promoter of *aly* gene as outgroup **(Figure 3.2.2)**. Expressed and non-expressed alleles of gene_090 are clustered separately with 72% bootstrap value and share unknown common ancestor. This suggest that expression of gene_090 could be evolved once during the course of evolution **(Figure 3.2.2)**.



```
aly     TTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATAT
360     TTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATAT
517     TTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATAT
307     TTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATAT
399     TTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATAT
304     TTGGGTCCCAATCCATAGGTGCACTGGAAGTATTTAATAATCTTCGCATATAACTAATAT
357     TTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAATAATCTTCGCATATAACTAATAT
ref     TTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAATAATCTTCGCATATAACTAATAT
        ****************** ***************** ***********************

aly     ATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTAT
360     ATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTAT
517     ATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTAT
307     ATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTAT
399     ATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTAT
304     ATCTTTATCTATATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTAT
357     ATCTTTATCTATATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTAT
ref     ATCTTTATCTATATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTAT
        ********** *************************************************

aly     TAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATA
360     TAAGAAAATCAGTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATA
517     TAACAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATA
307     TAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATA
399     TAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATA
304     TAAGACAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGGATAGTACTACATA
357     TAAGACAATCATATGA-TTTTTTGTGTTTCAGATACATCAAAAATGCATAGTACTACATA
ref     TAAGACAATCATATGA-TTTTTTGTGTTTCAGATACATCAAAAATGCATAGTACTACATA
        *** * *****  *** ****** ************** ******* *************

aly     -------CTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTAT
360     -------CTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTAT
517     -------CTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTAT
307     -------CTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTAT
399     -------CTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTAT
304     ATAATGGCTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGCTGAGCTAT
357     ATAATGGCTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGCTGAGCTAT
ref     ATAATGGCTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGCTGAGCTAT
               ***************************** ************** *** ****

aly     C-----CATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAA
360     C-----CATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAA
517     C-----CATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAA
307     C-----CATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAA
399     C-----CATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAA
304     CTATTATATATTTGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAA
357     CTATTATATATATGTGTTACCATTTAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAA
ref     CTATTATATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACAAGTAGAAA
        *      **** *******  *** *************************** *******
```

Figure 3.2.1- Expression profile and sequence alignment of *de novo* gene_090. A) Sequence alignment of expressing and non-expressing allele of gene_090. Gene_090 is expressed in RAL-307, 360, 399 and 517 lines (shown in red box) but not expressed in RAL- 304 and 357 lines (shown in blue box). aly represents the sequence of allele of gene_090 from *aly* chromosome which is also expressed (shown by red box), ref represents the reference genome sequence. Reverse complementary sequences were used for the alignment. SNPs and indels are highlighted by green colour. Rounded rectangular box highlights the TCE site. TSS shown in cyan colour and transcription progress by yellow arrow.

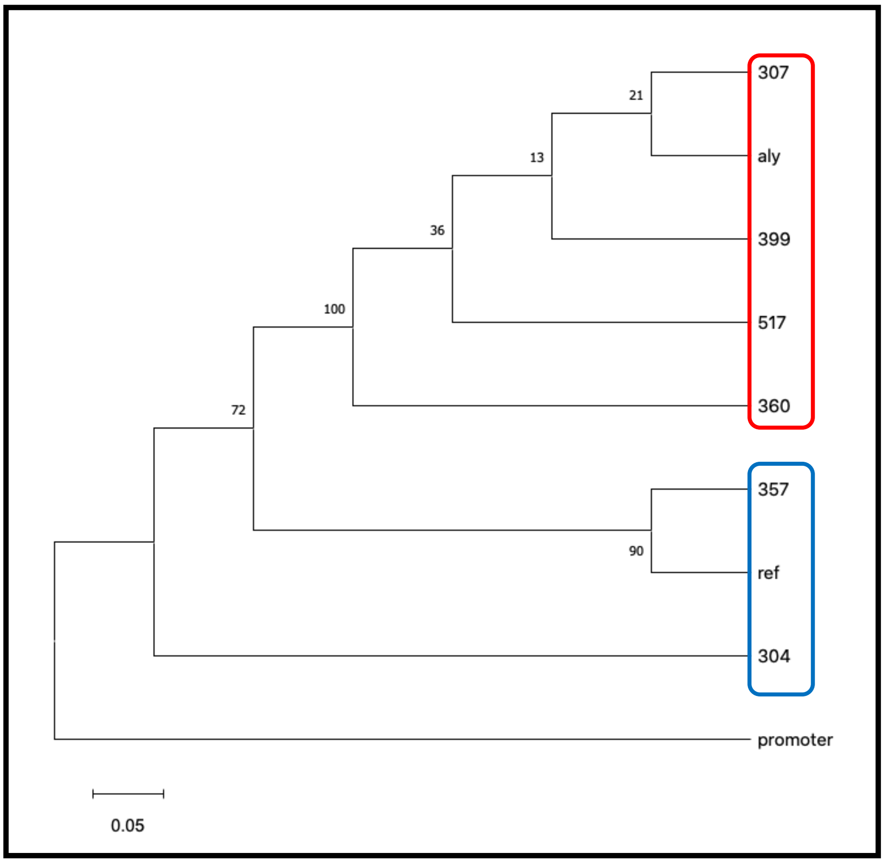


Figure 3.2.2:- A phylogenetic tree for gene_090. A phylogenetic tree showing expressed alleles are shown by red box and non-expressed alleles are shown by blue box cluster separately. The promoter is promoter of *aly* gene used as outgroup. The number represents the bootstrap values (%). The tree was generated using maximum likelihood statistical test with 1000 bootstrap iterations. The tree is drawn to scale, with branch lengths measured in the number of substitution per site. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. Scale represents the distance as percent divergence.

3.3. Cloning of attB site into reporter construct for site specific recombination in the genome

For genomic integration, pCaSpeR-AUG-β-Gal vector (Thummel, Boulet, and Lipshitz 1988) was modified by introducing the attB site. attB site was excised from pUAST attB plasmid using *Pst1* restriction enzyme **(Figure 3.3A&B)** and excised attB site was amplified using primer set flanking *Pst* 1 restriction enzyme site (see appendix 1). The amplified product was checked by agarose gel and ~300bp long product was observed **(Figure 3.3C)**, purified by Gel purification kit (QIAGEN), and sequenced by Sanger sequencing. The sequenced attB site was used for cloning into vector pCaSpeR-AUG-β-Gal-*djl*Δ52. A detailed map of this vector is given below showing the main features in

Figure 3.4.1. For attB site cloning into vector pCaSpeR-AUG-β-Gal, vector was digested with *Pst* 1 enzyme and attB site was cloned into the vector by Gibson assembly **(Figure 3.3E)**. Colony PCR was performed to confirm the clone and corresponding colonies was picked for Mini plasmid preparation **(Figure 3.3F)**. The cloned attB site was confirmed by Sanger sequencing. The attB site cloned vector was used for cloning of *de novo* gene's fragment. Detailed design and cloning of synthetic DNA of gene_090 is described in the material and methods chapter.

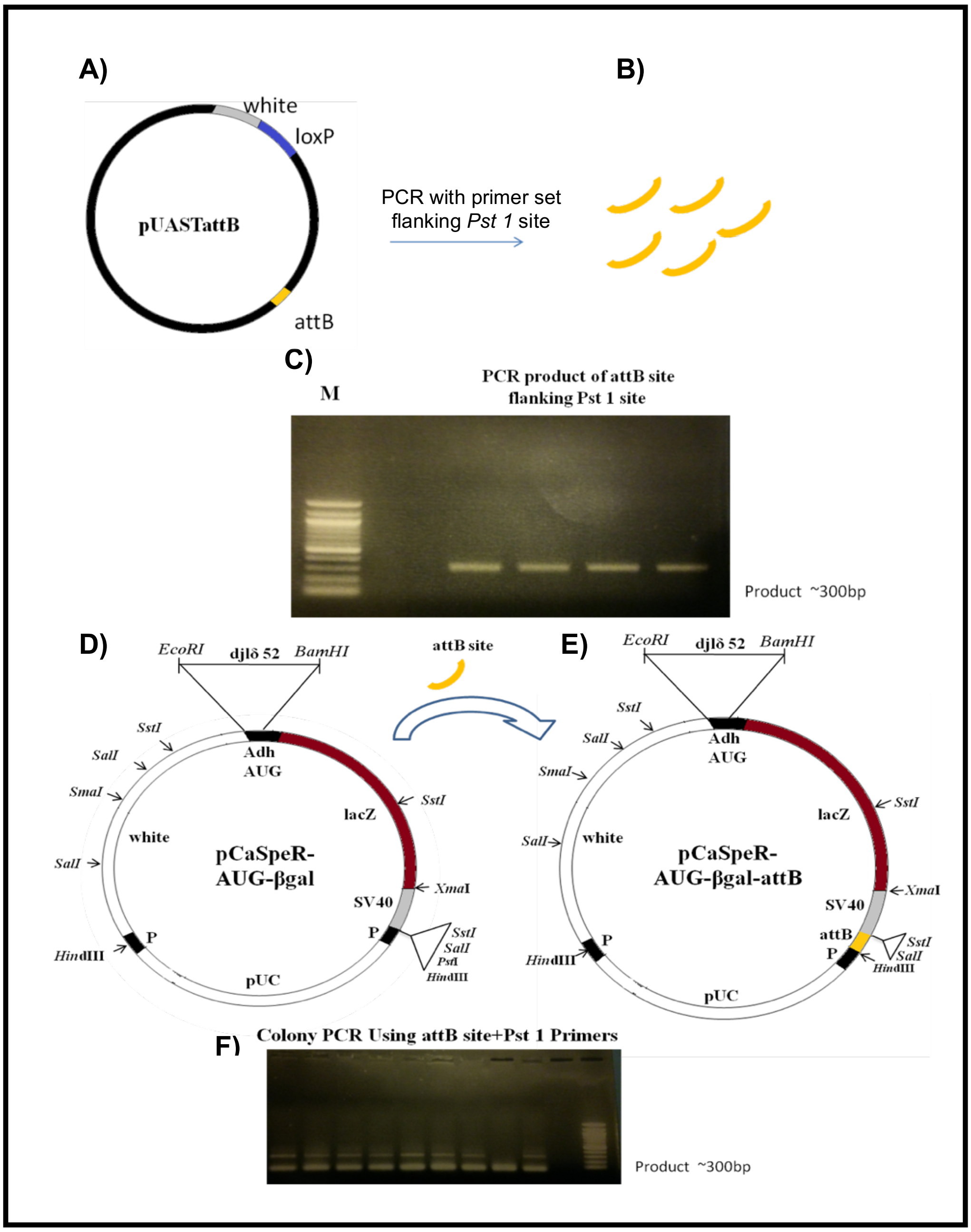


Figure 3.3:- Schematic representation showing generation of attB site and cloning into pCaSpeR-AUG-bGal vector. A) pUAST-attB plasmid with attB site, B) Excised attB site by Pst1 digestion, C) attB site amplification and gel electrophoresis of attB flanking *Pst* 1 site D) pCaSpeR-AUG-bGal digestion with Pst1 enzyme and E) Cloning of attB site in pCaSpeR-AUG-bGal vector. F) Colony PCR results confirm the insertion of attB site in vector.

3.4. Features of pCaSpeR-AUG-bGal-attB reporter construct

All synthetic DNA fragments for gene_090 were cloned into pCaSpeR-AUG-bGal-attB reporter construct **(Figure 3.4.1 & 3.4.2).** This vector lacks a minimal promoter and has a start codon to express reporter gene. All synthetic DNA fragments of gene_090 were cloned between *EcoR*I and *BamH*I restriction sites. MCSs were created by digestion of vector with *EcoR*I and *BamH*I to excise the *djl* fragment and insertion of polylinker but Gibson assembly cloning converted the *BamH*I site to a *Kpn*I site.

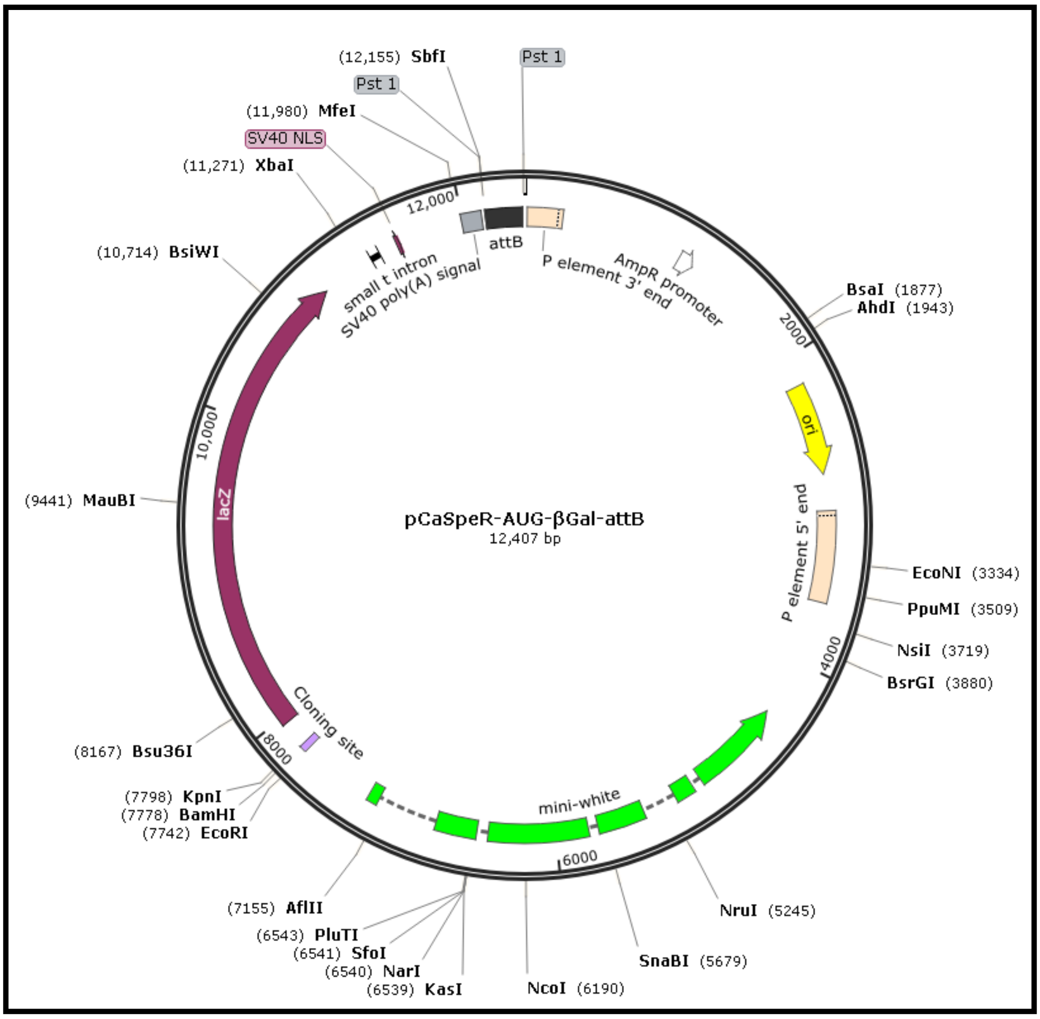


Figure 3.4.1:- Schematic representation of pCaSpeR-AUG-bGal-attB vector. Main features of this reporter construct are *white* gene (green) as a selection marker, multiple cloning sites (MCS) (purple) (it was created by removing the *djl* fragment and inserting polylinker) for cloning, 5' UTR from *adh* gene, *LacZ* gene coding sequence (red), 3' UTR from SV40 virus, and attB site (black) for site specific recombination in the genome, ori (yellow), ampicillin resistant etc. Size of the vector is about 12.5kb.

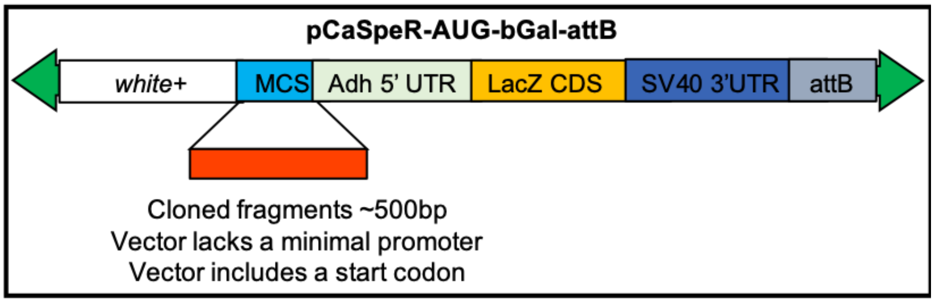


Figure 3.4.2:- Basic properties of pCaSpeR-AUG-bGal-attB reporter vector.

3.5. Reporter constructs comprising an expressed or non-expressed allele promoter recapitulate the endogenous expression pattern of gene_090

To investigate the mechanism that determines whether a specific *de novo* gene allele is expressed or not expressed, we tested the promoter region in reporter constructs. Based on previous reporter constructs this was likely to contain the genomic sequences of the target gene required for expression in testis (Thummel, Boulet, and Lipshitz 1988, White-Cooper 2012). Previously in our lab, it was discovered that expressed and non-expressed allele of gene_090 differs by 12 SNPs and two indels **(Figure 3.2.1)**. Also, phylogeny analysis of expressed and non-expressed allele of gene_090 showed that expression of gene_090 could be evolved once **(Figure 3.2.2)**. To test the promoter region of gene_090, a region comprising -302bp upstream of the transcription start site (TSS) and +133bp downstream of the TSS (with respect to *D. melanogaster* reference genome sequence) was selected **(Figure 3.2.1)**.

Two constructs were initially designed; 090_A contains the DNA sequence from an expressed allele of the gene while 090_B contains the sequence from a non-expressing allele of the gene **(Figure 3.5A&B)**. This experiment allowed me to test whether the region selected does indeed contain the sequences required for testis expression. It also allowed me to test whether the differences between the alleles in this region are responsible for the expression or non-expression.

Transgenic flies were generated by injecting the embryos from *yw;nos-φ31;attp40* flies with the vector and then the vector was inserted at the specific attP40 site via phi-C31 mediated recombination on the second chromosome. Then, transgenic flies were recovered by selecting for the w^+ marker.

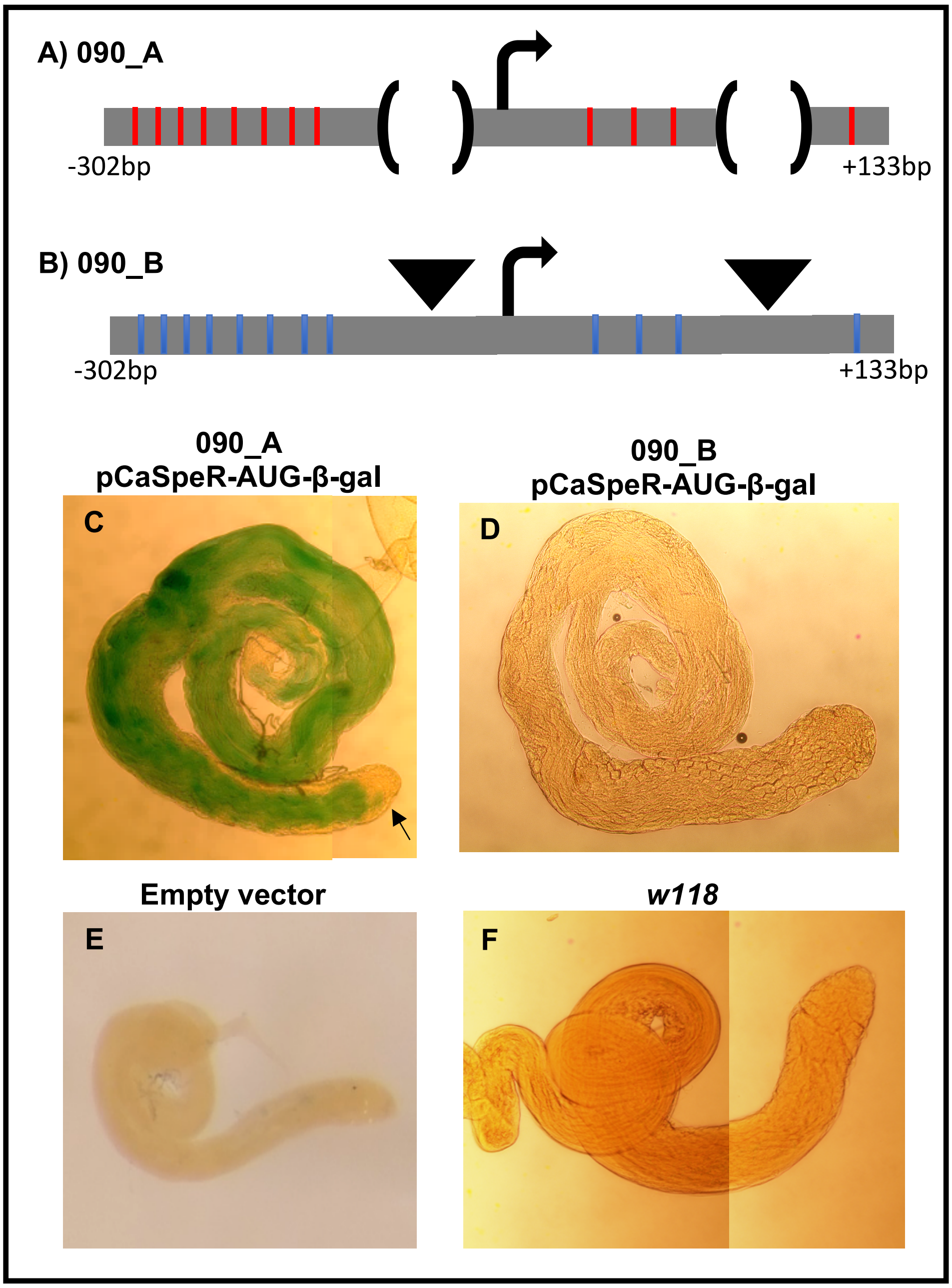


Figure 3.5:-Schematic representation of promoter sequences of gene_090 and X-gal staining in transgenic *D. melanogaster* testes . A) Expressed allele sequence with SNPs (in red) and indels i.e., deletion (as a bracket). B) Non-expressed allele sequence with SNPs (in blue) and indels i.e., insertions (as triangle). Differential interference contrast microscope images of whole testes. C) Testes of flies carrying the 090_A reporter construct of gene_090 showed beta-galactosidase activity staining throughout the testes except in germline stem cells (GSC) (arrow) D) Transgenic testes from 090_B construct showed no beta-galactosidase activity. E) Transgenic testes for empty vector showed no beta-galactosidase activity as wild type testes. F) Beta-galactosidase staining on testes from wild type flies *w*^{1118}*.* n=5 pairs of testis.

Importantly, beta-galactosidase activity staining on the testes from the transgenic line revealed that empty vector was not capable for driving expression of the reporter gene, hence showed no staining for the beta-galactosidase **(Figure 3.5E)**. It established that vector itself has no influence on reporter gene expression. Beta-galactosidase activity staining on the testes from the transgenic lines revealed that 090_A is robustly expressed in otherwise wild type testes and showed staining in the post meiotic cells of spermatogenesis such as spermatids but there was no obvious staining in germline stem cells, spermatogonia, spermatocytes, and somatic tissue **(Figure3.5C)**. An interesting observation from beta-galactosidase activity staining was that the reporter gene was translated in spermatids suggesting translational delay. In contrast, 090_B transgenic testes lacked any beta-galactosidase activity **(Figure3.5D),** and resembled testes from flies with an empty vector as well as w^{1118} flies **(Figure3.5E&F).** This reveals that -302bp to +132bp fragment indeed contain sequences capable of driving expression in the male germline. In addition, this shows that the difference between the expressing and non-expressing alleles can be attributed to sequence differences with this short yet important DNA sequence.

In light of these results, I hypothesized that it might be SNPs or indels (i.e., deletion or insertion as shown in figure 3.5A&B) which are important for expression. To address this hypothesis, a series of additional constructs were designed to test whether the SNPs or the indels, or both are responsible for the difference between expression and non-expression. In addition to the above, a previously characterised 10 bp sequence motif that regulates the translation in *Drosophila* testes known as the translation control element (TCE) also regulate the transcription in *Drosophila* testes (Katzenberger, R.J., et al., 2012). Gene_090 has a predicted TCE just downstream of TSS (**Figure 3.2.1)** and it would be interesting to check if it acts as a transcriptional regulator site. To test this hypothesis, a construct was designed with the non-expressed allele version of the TCE sequence, in an otherwise expressed allele background.

3.6. Reporter construct with mutation of predicted TCE show beta-galactosidase activity staining in testes

Intriguingly, bioinformatics analysis on sequences of expressed and non- expressed allele sequences predicted a TCE just downstream of the TSS. This TCE is also of interest because it differs between expressed and non-expressed allele and also a transcriptional regulator site **(Figure 3.2.1)** (Katzenberger et al. 2012). To determine the potential role of the predicted TCE, a construct was designed that contains the sequence from the expressed allele of gene_090 with single SNP which makes the sequence in the expressed allele a much better match to the TCE consensus than the non-expressed allele **(Figure 3.6A)**. The published TCE consensus sequence is *CTCAAAATTT*

(Katzenberger et al. 2012) whereas expressed allele has consensus of *ATCAAAATGC* and non-expressed allele has consensus of *ATAAAAATGC* **(Figure 3.2.1).**

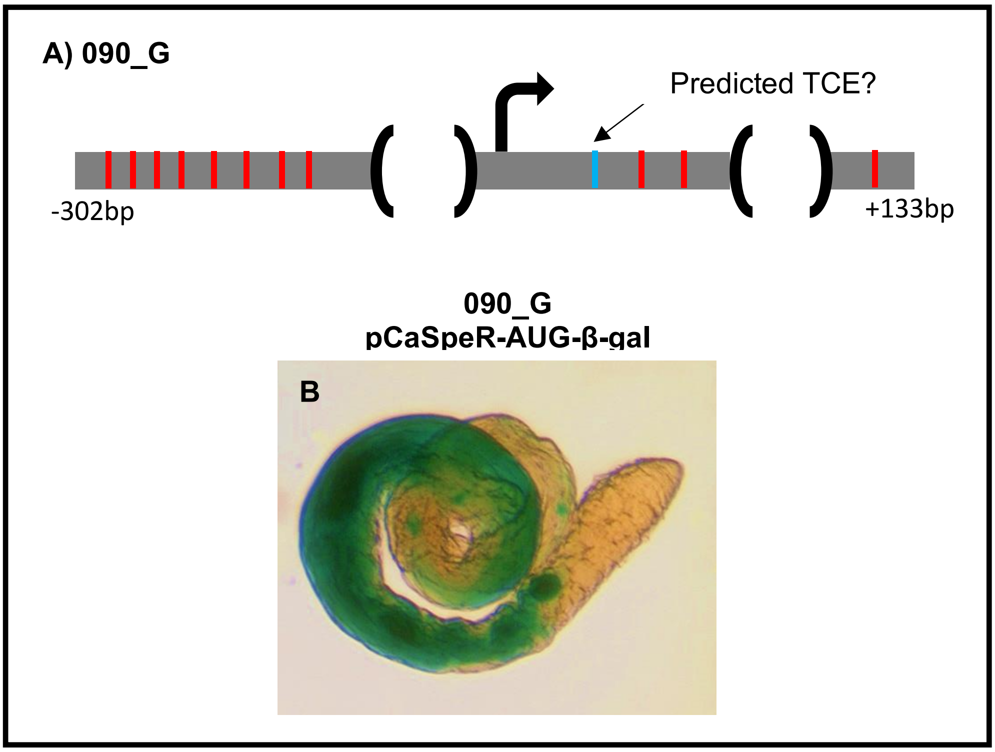


Figure 3.6- Representation of construct design of 090_G and beta-galactosidase staining activity in transgenic testes. A) Construct 090_G showing sequence from expressed allele with SNPs (shown by red colour), deletion (shown as bracket), predicted TCE (shown in blue colour), and TSS (shown as an arrow). B) X-gal staining on transgenic testes for construct 090_G showing beta-galactosidase activity staining in spermatids towards the more basal region of the testis but primary spermatocytes towards the tip of the testis were not stained. n=5 pairs of testis.

Transgenic testes of this construct showed the strong beta-galactosidase activity staining in spermatids towards the more basal region of the testis but primary spermatocytes towards the tip of the testis and GSC were not stained (**Figure 3.6B**). This result suggests that predicted TCE does not act transcriptional control site. Beta-galactosidase activity staining was a bit weaker, suggesting that TCE maybe enhance translation of reporter gene in testis. This promoter sequence in this construct behave the same as promoter sequence cloned into construct 090_A **(Figure 3.5A)**.

3.7. Only indels are important for expression of gene_090 not SNPs

It was established that promoter sequences of expressed and non-expressed allele of gene_090 showed expression and no expression respectively. The region flanking the TSS used in reporter constructs 090_A and 090_B differs by 12 SNPs and two indels. Now, I wanted to test if SNPs or indels or both are important for the expression of gene_090. In order to address this question, I designed two constructs, the first construct 090_C comprises the sequence from the expressed allele but both the indels i.e., insertions from non-expressed allele (**Figure 3.7A**). The second construct 090_D has the sequence from non-expressed allele but both the indels i.e., deletions (7 bp and 5 bp each) from the expressed allele (**Figure 3.7B**).

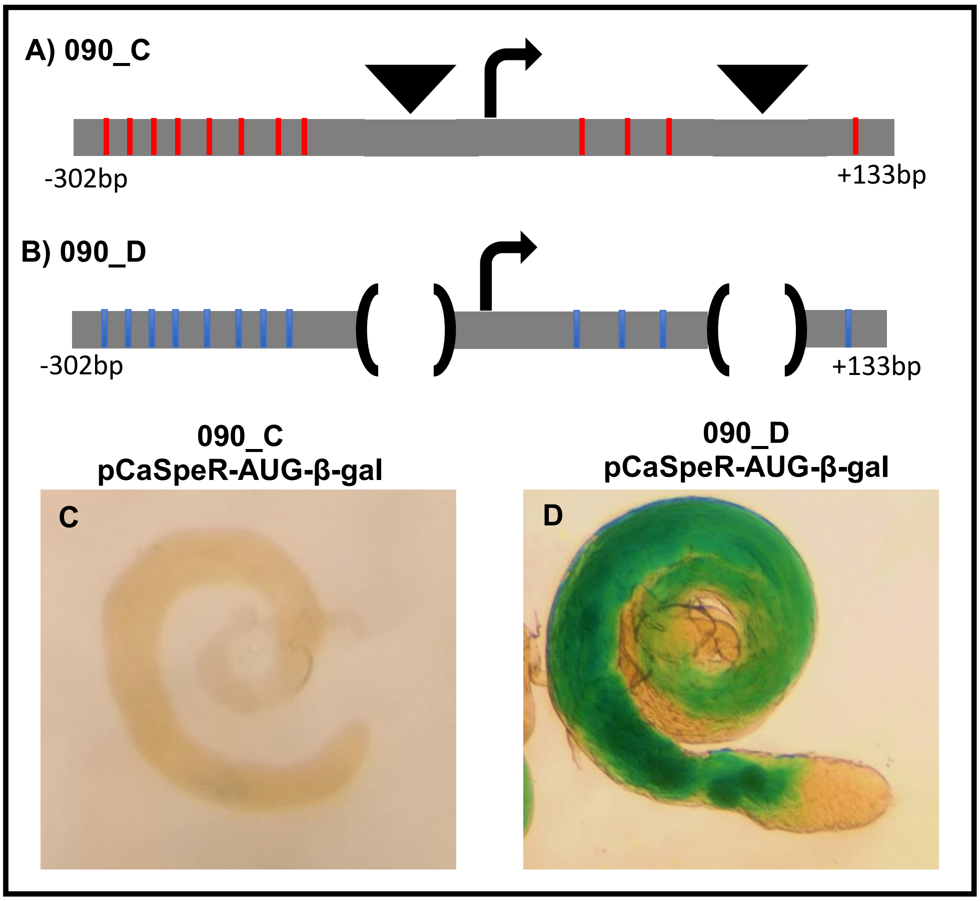


Figure 3.7:- Construct design and beta-galactosidase activity staining on transgenic lines- 090_C & 090_D. A) Schematic design of construct 090_C with SNPs are highlighted in red colour and both the indels i.e., insertion are shown by triangles. B) 090_D construct design showing SNPs in blue colour and both indels i.e., deletions showed by brackets. C) There was no beta-galactosidase activity staining in testes for transgenic flies for 090_C. D) Testis from the transgenic lines for 090_D showed strong expression. n=5 pairs of testis

Testes from flies with the construct 090_C with SNPs from expressed allele and indels (insertions) from non-expressed allele showed no X-gal staining (**Figure 3.7C**) at all whereas the second construct 090_D with SNPs from non-expressed allele and indels (deletions) from the expressed allele showed strong in spermatids but no evident staining in GSC, spermatogonia, spermatocytes, and somatic tissues (**Figure 3.7D**). These results suggest that SNPs are not important but indels i.e. deletions are important for expression of reporter gene.

3.8. 7bp Indel-I i.e. deletion is necessary for expression of gene_090

Earlier results showed that indels are important for expression of reporter gene in testis. Now, I wanted to check if both indels are important or whether the transcriptional activity could be attributed to just one of the indels. To test the necessity of the indels, I designed two more constructs. The first construct 090_E contain all the SNPs and indel-II (deletion) from expressed allele and indel-I (insertion) from non-expressed allele (**Figure 3.8A**). The second construct 090_F comprises of a sequence with all the SNPs and the indel-I (deletion) from expressed allele and the indel-II (insertion) non-expressed allele (**Figure 3.8B**). Transgenic testes for construct 090_E showed no beta-galactosidase staining (**Figure 3.8C**) indistinguishable from the negative control **(Figure 3.5E)** whereas construct 090_F showed beta-galactosidase staining in testes (**Figure 3.8D**) as robust as construct 090_A **(Figure 3.5C)**.

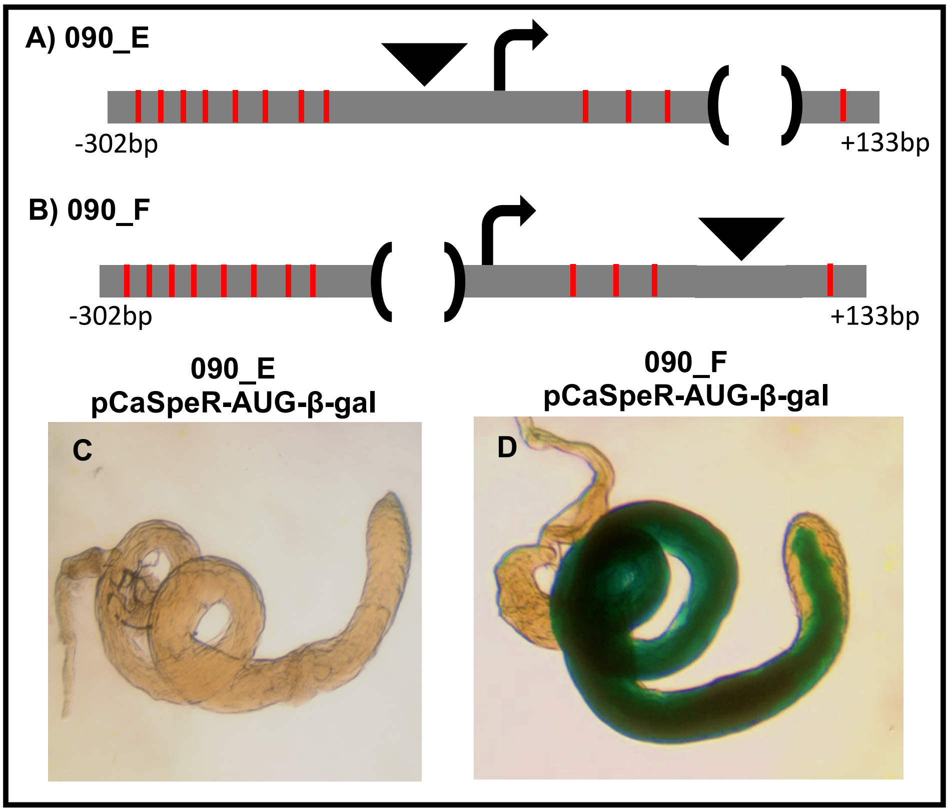


Figure 3.8- Schematic representation of construct designing and X-gal staining for beta-galactosidase activity in transgenic testes for 090_E & 090_F. A) Construct 090_E has SNPs from expressing allele showed in red colour, indel-I (insertion) by a triangle, and indel-II (deletion) by a bracket. B) Construct 090_F has SNPs from expressed allele highlighted in red colour, indel-I (deletion) shown as a bracket, and indel-II (insertion) shown as a triangle. C) Beta-galactosidase activity staining on transgenic testes from transgenic line for 090_E has no staining at all. D) Beta-galactosidase activity staining on testes from transgenic line for 090_F showed robust expression in testis. n=5 pairs of testis.

3.9. 7bp Indel-I i.e., deletion is necessary and sufficient to convert a natural non-expressing allele into an expressing allele

Having established that the 7bp deletion at the indel-I site is necessary for expression, now I wanted to test if this 7bp indel-I is sufficient for expression in testes. To test this hypothesis, I designed a construct with indel-I (deletion) from the expressed allele and SNPs and indel-II from the non-expressed allele (**Figure 3.9A**). Testes from transgenic flies were dissected and used for the x-gal staining. Excitingly, transgenic testes showed strong beta-galactosidase staining in post meiotic cells of spermatogenesis in testes (**Figure 3.9B**) and revealed that 7bp indel-I (deletion) is also sufficient for expression in testes.

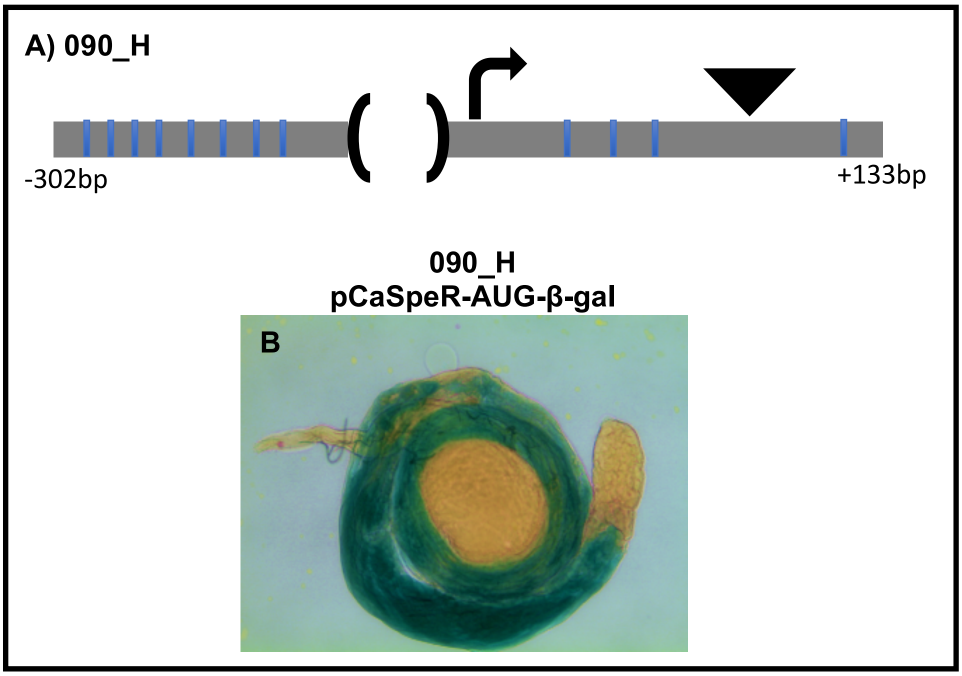


Figure 3.9:- Representation of 090_H construct design and X-gal staining in testes. A) The construct 090_H is composed of SNPs (by blue colour) and indel-I (by bracket), and indel-II. B) Beta-galactosidase activity staining on testes from transgenic lines for 090_H showing solid staining in testes.

Altogether, these results suggest that this specific 7 bp deletion is necessary and sufficient to convert a naturally occurring non-expressing allele into an expressing allele.

3.10. What is the mechanism of regulation?

Based on the results so far, I proposed three models that could be possible mechanisms of expression regulation of gene_090 in *Drosophila* testes.

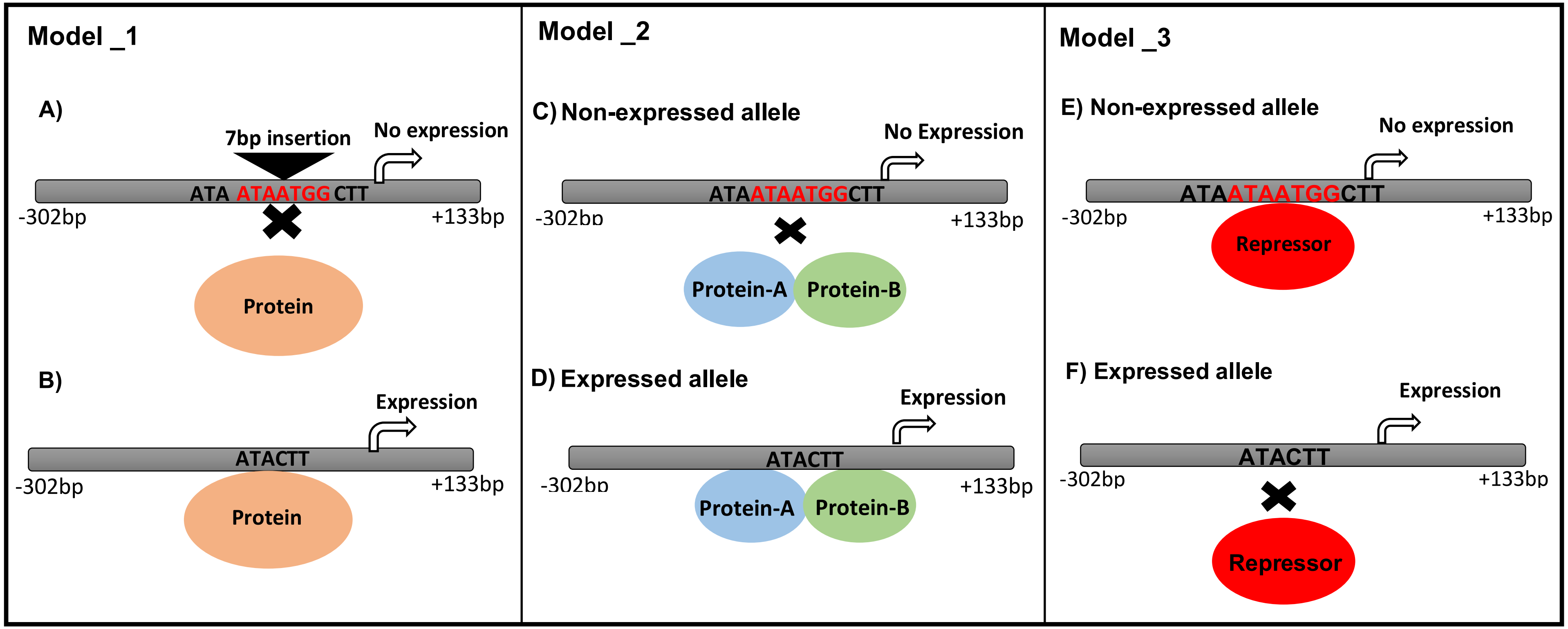


Figure 3.10- Proposed models describing possible mechanism of gene expression regulation. A&B) Model-1 showing when we have insertion of 7bp (shown by triangle) there no expression and when we have a deletion there is expression. Potential protein responsible for expression shown in pink. C, D, &E) Model-2 explaining another possible mechanism of gene expression regulation in testes where more than one protein is involved in expression regulation and these protein potentially bind to flanking sites of indel-I (deletion). F, G, &H) Model-3 describing repressor binding site and repressor molecule binding to this site regulates gene expression in testes.

The first model suggests that in non-expressing allele 7bp insertion (**ATAATGG**) (shown in blue box in figure 3.9A) just upstream of TSS cause disruption of regulatory protein binding site and hence gives no expression (**Figure 3.10A**). Deletion of this 7bp in an expressing allele possibly creates a new binding site by bringing flanking sites [**ATA()CTT**] together and allows regulatory protein to bind to the promoter and drives gene expression in testes (**Figure 3.10B**). The second model suggests there might be more than one protein involved in expression regulation and these two proteins bind to sites flanking the 7bp deletion (**Figure 3.10C&D**) and regulate the gene_090 expression. The third possible model suggests that 7bp insertion in non-expressing allele is a repressor binding site, which allows repressor molecule to bind to this site and thus prevent the gene expression (**Figure 3.10E&F**).

3.11. 7 bp deletion in expressing allele do not create a new binding site for any regulatory protein/proteins

It was established that indel-I i.e., deletion is necessary and sufficient for expression of gene_090. Now I wanted to address 1) if indel-I i.e., deletion is making any new binding site in expressed allele background. 2) If Indel-I i.e., insertion disrupts this binding site in non-expressed allele background. To test this hypothesis, I created several constructs carrying sequences from immediately upstream and downstream of indel-I in expressing and non-expressing allele background. With these constructs, we could understand if the flanking sites of indel-I are creating any new binding site with deletion. Using 7 bp sequence from upstream as well as downstream would allow us to keep the sequences length same as non-expressing allele.

3.11.1. 7 bp duplication of the sequence from upstream of the indel-I site do not create a new binding site

In construct 090_I, 7 bp sequence (**CTACATA**) is inserted immediately upstream of indel-I at indel-I site in expressed allele background (**Figure 3.11.1B**) whereas 090_J has the same insertion (**CTACATA**) at the indel site-I but in non-expressed allele background (**Figure 3.11.1D**). Transgenic flies for these constructs were generated and testes were stained for beta-galactosidase activity. Interestingly, both the constructs 090_I and 090_J showed no beta-galactosidase activity in testes (**Figure 3.11.1F&H**), thus suggesting insertion of 7bp sequence from upstream of indel-I at indel-I site is not potentially creating any new binding site for regulatory protein/proteins for gene expression regulation.

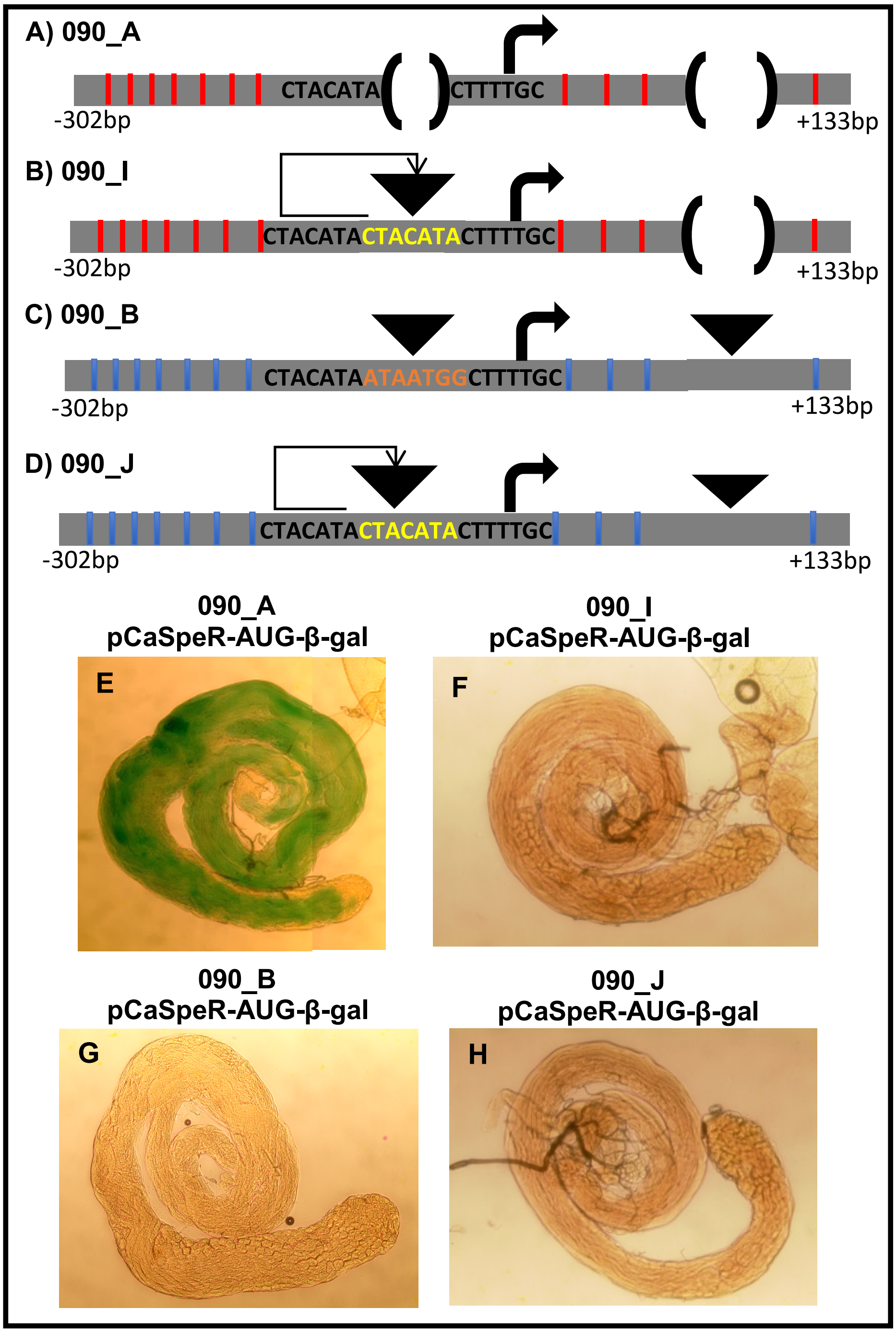


Figure 3.11.1:- Schematic representation of 090_I & 090_J constructs designing and X-gal staining for beta-galactosidase activity. A&B) Construct 090_A and design of 090_I involves expressed allele sequence with SNPs (shown in red colour) and insertion of 7bp at indel-I site (shown by inverted arrow and sequence yellow colour), and indel-II shown as a bracket. B&D) Construct 090_B and design of construct 090_J has non-expressed allele sequence with SNPs (in blue colour), insertion of 7bp at indel-I site (shown by inverted arrow and sequence yellow colour), and indel-II shown as a triangle. E) Beta-galactosidase activity staining on transgenic testes for construct 090_A, and F) for construct 090_I, G) for construct 090_B, and H) for construct 090_J. n=5 pairs of testis.

Staining of these transgenic tests resemble the staining of empty vector (shown in Figure 3.5E). In addition to these results, I also wanted to test whether insertion of 7bp sequence (**CTTTTGC**) from downstream of the indel-I at indel-I creates any new binding site.

3.11.2. 7 bp duplication of the sequence from downstream of the indel-I site also do not create a new binding site

To investigate if insertion of sequence immediately downstream of indel-I at indel-I site create a binding site or not, I designed two construct and the first construct 090_K contain insertion of 7 bp sequence from downstream of indel-I at indel-I site in expressed allele background (**Figure 3.11.2B**). The second construct 090_L contain insertion of 7bp sequence from downstream of indel-I at indel-I site in non-expressed allele background (**Figure 3.11.2D**).

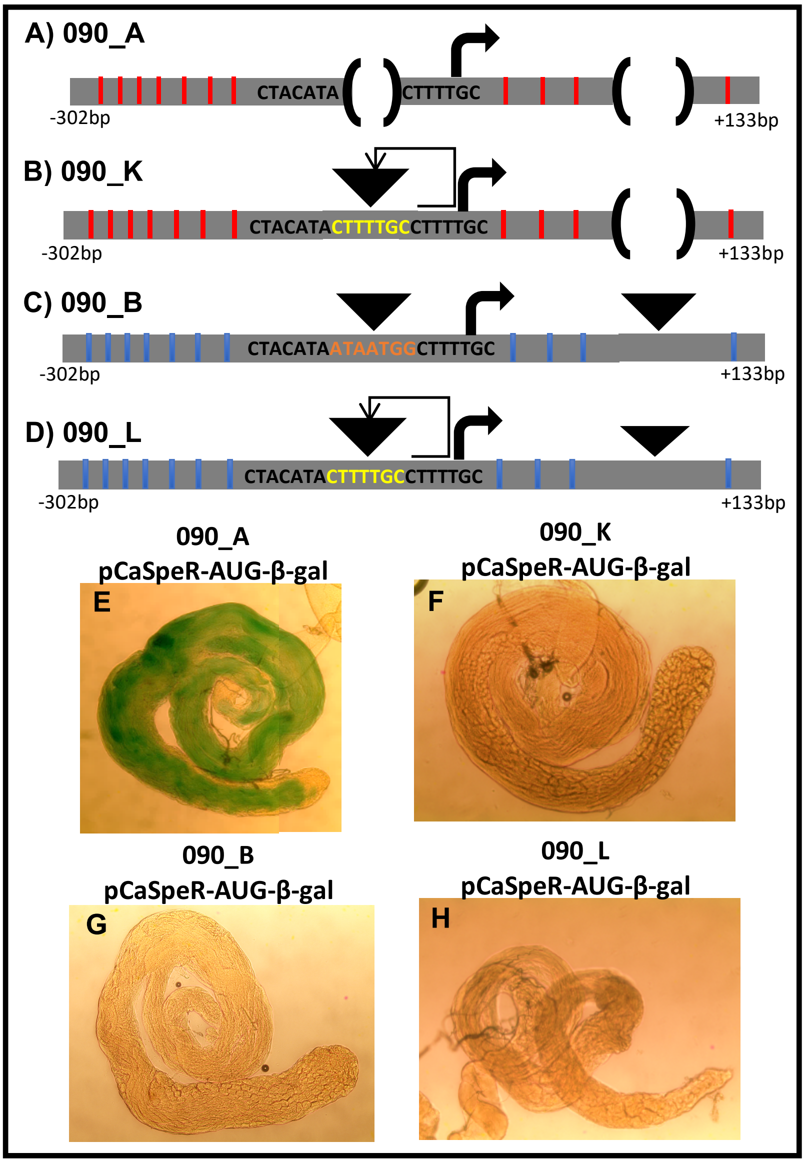


Figure 3.11.2:- Representation of 090_K & 090_L constructs design and x-gal staining on testes from transgenic lines. A&B) Construct 090_A and construct design for 090_K with SNPs (in red colour), insertion at indel-I site (shown by inverted arrow and sequence with yellow colour), and indel-II (shown as a bracket). B&D) Construct 090_B and design of construct 090_L has SNPs (blue colour), insertion at indel-I site (shown by inverted arrow and sequence with yellow colour), and indel-II (shown as a triangle). E&F) Showing the beta-galactosidase activity staining on testes from transgenic lines for construct 090_A and 090_K respectively and G&H) for construct 090_B and 090_L respectively. n=5 pairs of testis.

X-gal staining on transgenic testes for these both constructs showed no beta-galactosidase activity staining (**Figure 3.11.2F&H**), hence suggesting that duplication of sequence from downstream of indel-I at indel-I site does not create a binding site for regulatory protein/s. Now it was proven that 7bp insertion of sequence from upstream or downstream of indel-I at indel-I site does not create any binding site or any site created longer than the region tested with these short duplications and suggests that model_2 (**Figure 3. 10C&D**) is possible mechanism of gene regulation in which more than one protein is involved in gene expression regulation of gene_090. Furthermore, I wanted to test if there is a repressor binding site being made at indel-I (insertion) site and for that I devised two constructs with scramble and inverted version of indel-I (insertion) from non-expressed allele in expressed allele background.

3.12. Scramble and inverted version of indel-I i.e., insertion also do not create a new binding site

To test these hypothesizes, two construct were designed and both constructs contain sequence from expressed allele but indel-I (insertion) is from non-expressed. For construct 090_M, indel-I (insertion) was scrambled **(TAGAGAT)** and inserted at indel-I site (**Figure 3.12A**) and for construct 090_N indel-I (insertion) was inverted (**CCATTAT**) and inserted at indel-I site (**Figure 3.12B**).

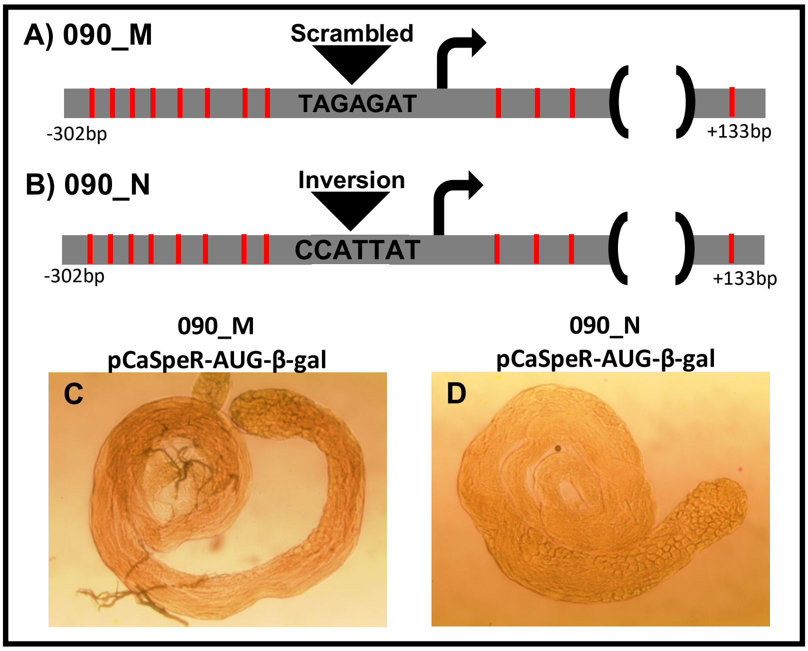


Figure 3.12:- Construction of 090_M and 090_N construct and x-gal staining on transgenic testes. A) Construct 090_M has expressed allele sequence with SNPs (shown by red colour), scrambled indel-I at indel-I site (shown by a triangle), and indel-II (as a bracket). B) Construct 090_N with SNPs (shown by red colour), inverted version of indel-I (shown by triangle), and indel-II (as a bracket). C&D) Beta-galactosidase activity staining on testes from transgenic lines for construct 090_M and 090_N respectively. n=5 pairs of testis.

Testes from transgenic lines for these both constructs showed no beta-galactosidase activity staining (**Figure 3.12C&D**). These results are consistent with the possible mechanism of gene expression regulation could be model-2 (**Figure 3. 10C&D**) and suggest that the insertion does not generate a sequence that recruits a repressor.

Overall these results suggest that we are not creating any binding site with the deletion but that the regions flanking insertion could be important for expression. With an insertion at indel-I, a potential binding site for regulatory proteins could disrupt therefore no expression. Altogether this suggests that model-2 is most probable mechanism of gene expression where two proteins might bind to the flanking sites of the indel-I i.e., deletion. Based on these results, I designed two more constructs to check if the indel-I (insertion) is a repressor binding site or not.

3.13. Alteration of indel-I i.e., insertion position in promoter region does give expression in primary spermatocytes

To address if moving the indel-I (insertion) from its original position could alter the expression or not, I designed two constructs. First construct 090_O contain sequence from expressed allele but the sequence of indel-I (insertion) moved 20 bp (two DNA helix turn) upstream of its original position (**Figure 3.13A**). Second construct 090_P contain the same sequence as construct 090_O but the indel-I (insertion) is moved 15 bp (1.5 DNA helix turn) upstream of its original position (**Figure 3.13B**). The notion of two different constructs with 1.5 and 2 DNA helix turn was to allow enough space to be available for regulatory proteins to bind at indel-I site.

Transgenic lines were generated for these two constructs and testes were stained for beta-galactosidase activity. Interestingly, beta-galactosidase activity staining revealed that construct 090_O is expressed in primary spermatocytes (**Figure 3.13C**) whereas other constructs were expressed in later stage of spermatogenesis and showed staining in late spermatids **(Figure 3.5C)**. On the other hand, transgenic testes for 090_P showed no staining at all (**Figure 3.13D**). These results suggest that insertion of indel-I i.e., insertion at 20 bp upstream of its original position altered the transcription of reporter gene and suggest it could be transcribed in spermatocytes or premeiotic stages of spermatogenesis. 15 bp upstream insertion lacked any beta-galactosidase activity, and resembled the testes from flies with empty vector inserted **(Figure 3.5E)**.

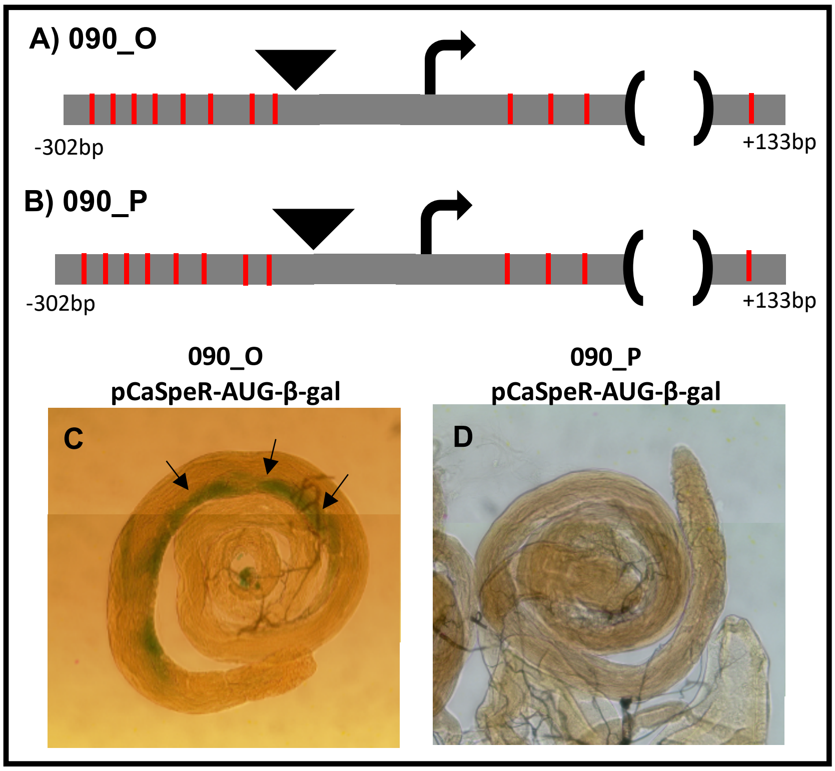


Figure 3.13:- Design of construct 090_O and 090_P and beta-galactosidase activity staining. A) Design of construct 090_O with expressed allele sequence showing 20bp upstream insertion of indel-I (shown by triangle), SNPs (shown by red colour), and indel-II (as a bracket). B) Construct 090_P has expressed allele background with SNPs (shown by red colour), 15bp upstream insertion of indel-I, and indel-II (as a bracket). C&D) Showing beta-galactosidase activity staining of the testes from transgenic lines for construct 090_O and 090_P respectively. n=5 pairs of testis.

3.14. Mutation of nucleotides flanking the deletion site in expressed allele give no expression in *Drosophila* testes

A deletion in the expressed allele allows flanking sites to come together and gives the expression. Here I wanted to test if switching the adjoining base pairs of deletion are implicated expression in testes. To address this hypothesis we devised a construct with two nucleotides are switched around deletion site in expressed allele background **(Figure 3.14B)**. Transgenic flies were generated for this construct and testes for this transgene were subjected to beta-galactosidase activity staining. There was no visible beta-galactosidase activity staining observed in testes **(Figure 3.14D)** and testes resembled the negative controls **(Figure 3.5E).** Thus it shows that a site created by deletion and adjoining nucleotides of the flanking sites of deletion are important for driving expression in *Drosophila* testes. It also shows that specificity of the sequence around deletion in expressed allele is important for expression.

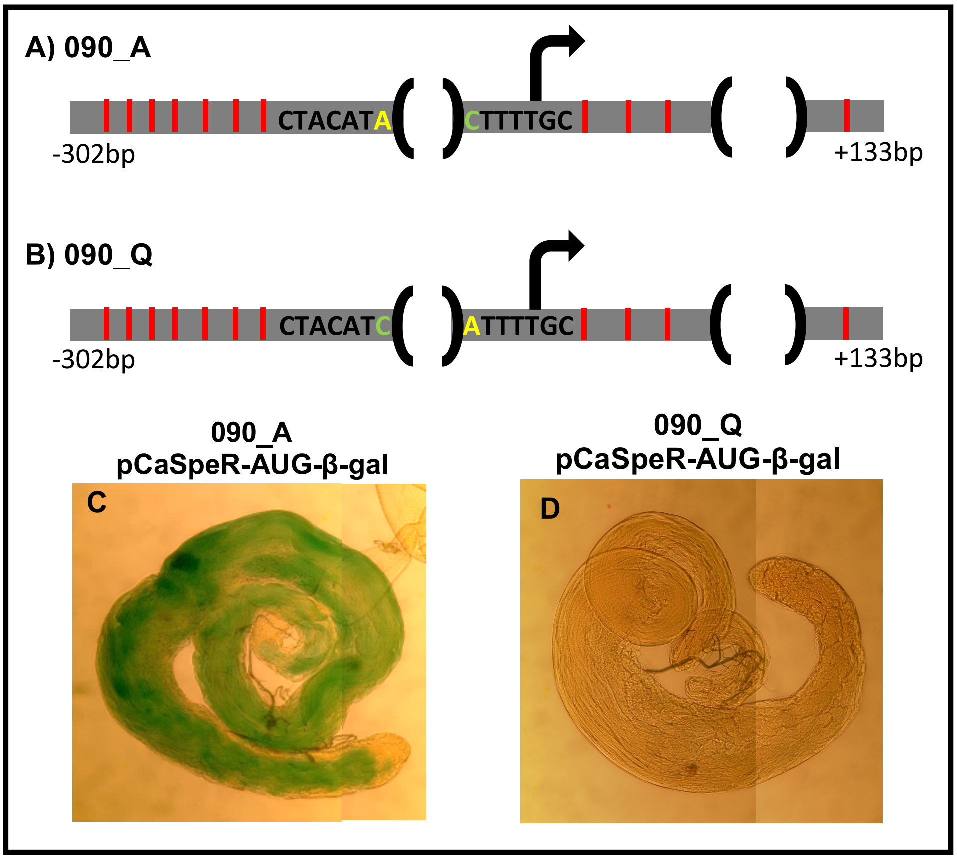


Figure 3.14:- Switching the flanking nucleotide of deletion does not give expression. A) Construct 090_A showing the flanking bases of deletion (highlighted by green "C" and yellow "A" colour). B) Construct 090_Q design showing SNPs by red colour and indels from the expressed allele by brackets and two nucleotide switched shown by green "C" and yellow "A" colour. C&D) Beta-galactosidase activity staining on construct 090_A and 090_Q respectively. n=5 pairs of testis.

3.15. Summary

Gene_090 is shown to be very interesting gene to study with small stretch of DNA has the potential to gain the expression in testes. SNPs present in promoter region of this gene has no effect on its expression whereas indel-I (deletion) created in expressed allele plays a vital role in gene expression regulation. 7 bp deletion in expressed allele is necessary as well as sufficient to convert naturally occurring non-expressing allele into expressing allele. Our results suggest that deletion in the expressed allele does not create any binding site for repressor protein per se but the mutation of the two base pairs of the flanking sites of the deletion suggest that the sequence created by the deletion is critical, as well as the spacing of other flanking sequences are important for expression. Our results also shows that 20 bp insertion of indel-I (insertion) from its original site in expressed allele background gives expression in primary spermatocytes.

Chapter 4. Results

Molecular analysis of mutations leading to expression of gene_074

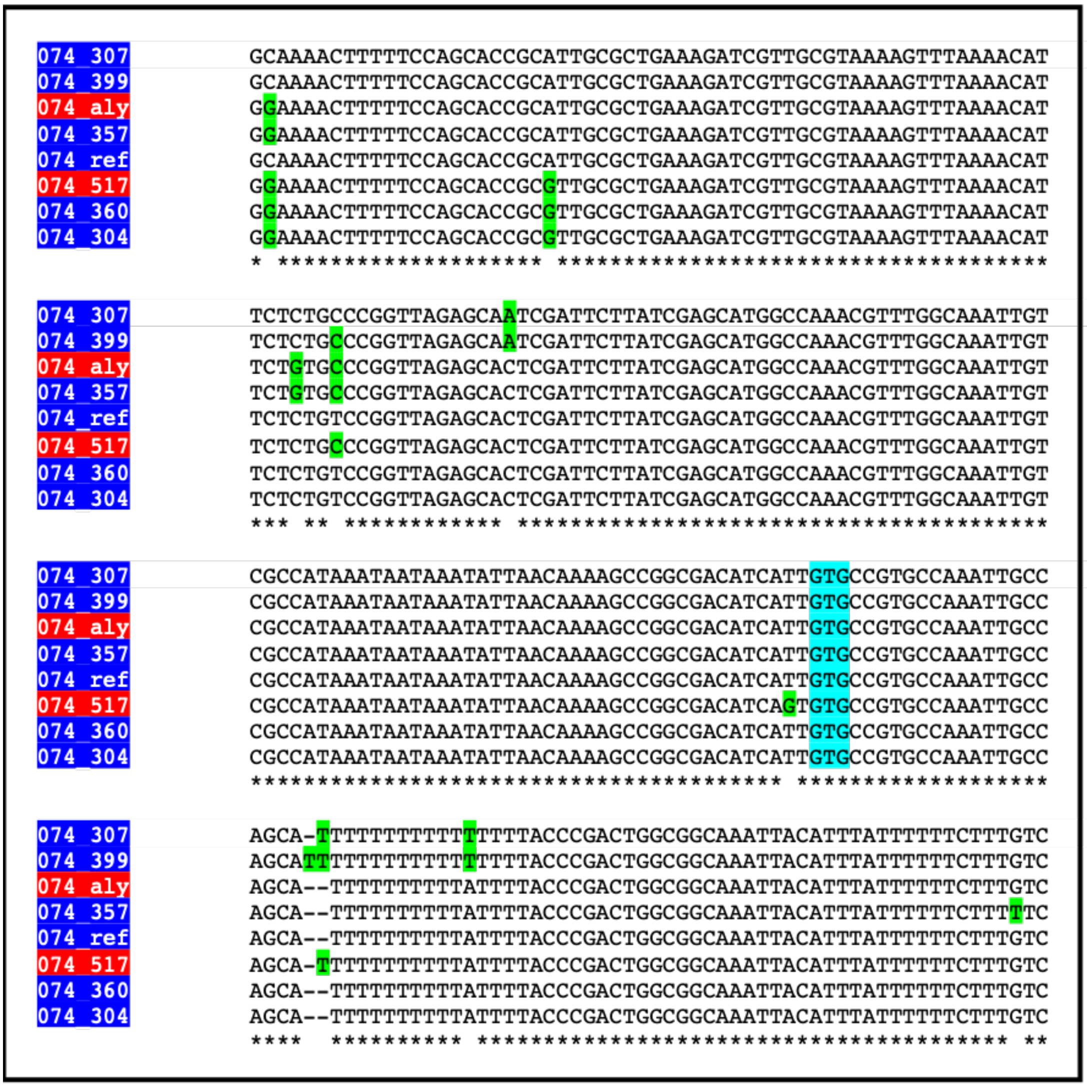
4.1. Abstract

De novo genes are new in populations and studying these genes would be interesting to understand the biology behind the evolution and gene expression regulation. One such interesting gene_074 (Zhao et al. 2014) has nucleotide polymorphisms in the promoter region which does not correlate with its expression. However, this gene is interesting to study because two expressing alleles differ by few SNPs and one indel whereas expressing allele and non-expressing allele cannot be differentiated by SNPs or indels **(Figure 4.2.1)**. This gene is also important to study because its expression is independent of tMAC complex (HW-C unpublished data). To understand the mechanism of gene expression regulation of gene_074, we designed several constructs and these constructs revealed that SNPs or indels are not important for the expression and additional upstream or downstream sequences from the TSS might be important for the expression.

4.2. The molecular properties and evolution of gene_074

Gene_074 is segregating *de novo* gene whose expression is not fixed in the population yet. This gene is expressed in line RAL-517 (159.80 FPKM) but not in other lines RAL-304 (0.26 FPKM), 307 (0.00 FPKM), 357 (0.00 FPKM), 360 (0.00 FPKM), 399 (0.09 FPKM), and the allele on the *aly1* chromosome. Expressing allele RAL-517 and non-expressing alleles of gene_074 differ by only 1 SNP **(Figure 4.2.1)**. In addition to above, two different expressing allele 074_517 and 074_aly1 can be differentiated by presence of 3 SNPs and 1 indel **(Figure 4.2.1)**.

A phylogenetic tree was generated using promoter regions of the gene_074 and promoter of *aly* gene as outgroup showed that both expressing alleles 074_517 and 074_aly1 fall in different clusters. Three non-expressing alleles 074_399, 307, and ref clusters together with bootstrap value- 57%. Another group of non-expressing alleles 074_360 and 304 cluster together with bootstrap value-66% and this cluster could be sharing common ancestor with 074_517 allele with 57% bootstrap value. The phylogenetic tress suggest that gene_074 could be evolved twice **(Figure 4.2.2)**.



```
074_307             GCAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACAT
074_399             GCAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACAT
074_aly             GGAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACAT
074_357             GGAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACAT
074_ref             GCAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACAT
074_517             GGAAAACTTTTTCCAGCACCGCGTTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACAT
074_360             GGAAAACTTTTTCCAGCACCGCGTTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACAT
074_304             GGAAAACTTTTTCCAGCACCGCGTTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACAT
                    * ******************** *************************************

074_307             TCTCTGCCCGGTTAGAGCAATCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT
074_399             TCTCTGCCCGGTTAGAGCAATCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT
074_aly             TCTGTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT
074_357             TCTGTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT
074_ref             TCTCTGTCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT
074_517             TCTCTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT
074_360             TCTCTGTCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT
074_304             TCTCTGTCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT
                    *** ** ************ ****************************************

074_307             CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCC
074_399             CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCC
074_aly             CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCC
074_357             CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCC
074_ref             CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCC
074_517             CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCAGTGTGCCGTGCCAAATTGCC
074_360             CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCC
074_304             CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCC
                    **************************************** *******************

074_307             AGCA-TTTTTTTTTTTTTTTACCCGACTGGCGGCAAATTACATTTATTTTTTTCTTTGTC
074_399             AGCATTTTTTTTTTTTTTTTACCCGACTGGCGGCAAATTACATTTATTTTTTTCTTTGTC
074_aly             AGCA--TTTTTTTTTTATTTACCCGACTGGCGGCAAATTACATTTATTTTTTTCTTTGTC
074_357             AGCA--TTTTTTTTTTATTTACCCGACTGGCGGCAAATTACATTTATTTTTTTCTTTTTC
074_ref             AGCA--TTTTTTTTTTATTTACCCGACTGGCGGCAAATTACATTTATTTTTTTCTTTGTC
074_517             AGCA-TTTTTTTTTTTATTTACCCGACTGGCGGCAAATTACATTTATTTTTTTCTTTGTC
074_360             AGCA--TTTTTTTTTTATTTACCCGACTGGCGGCAAATTACATTTATTTTTTTCTTTGTC
074_304             AGCA--TTTTTTTTTTATTTACCCGACTGGCGGCAAATTACATTTATTTTTTTCTTTGTC
                    ****  ********** **************************************** **
```

Figure 4.2.1:- Gene_074 expression profile and promoter sequence alignment. Gene_074 is expressed in line RAL-517 but not expressed in other lines RAL-304,307, 357, 360, and 399. Sequence alignment between expressing (highlighted in red color) and non-expressing (highlighted in blue color) alleles of gene_074 showing the predicted TSS (shown by cyan color) and several SNPs (highlighted in green color) upstream and downstream of TSS. 074_aly1 is the allele of gene_074 which is from aly chromosome.

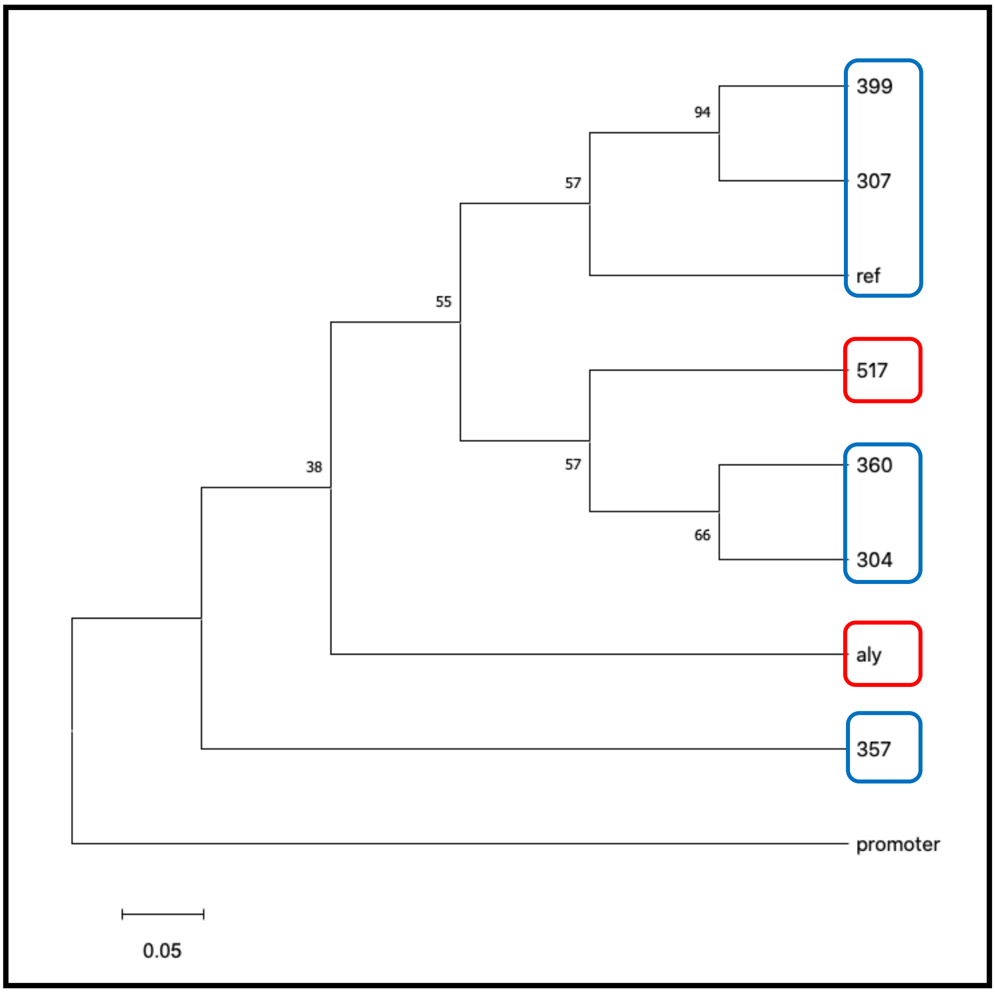
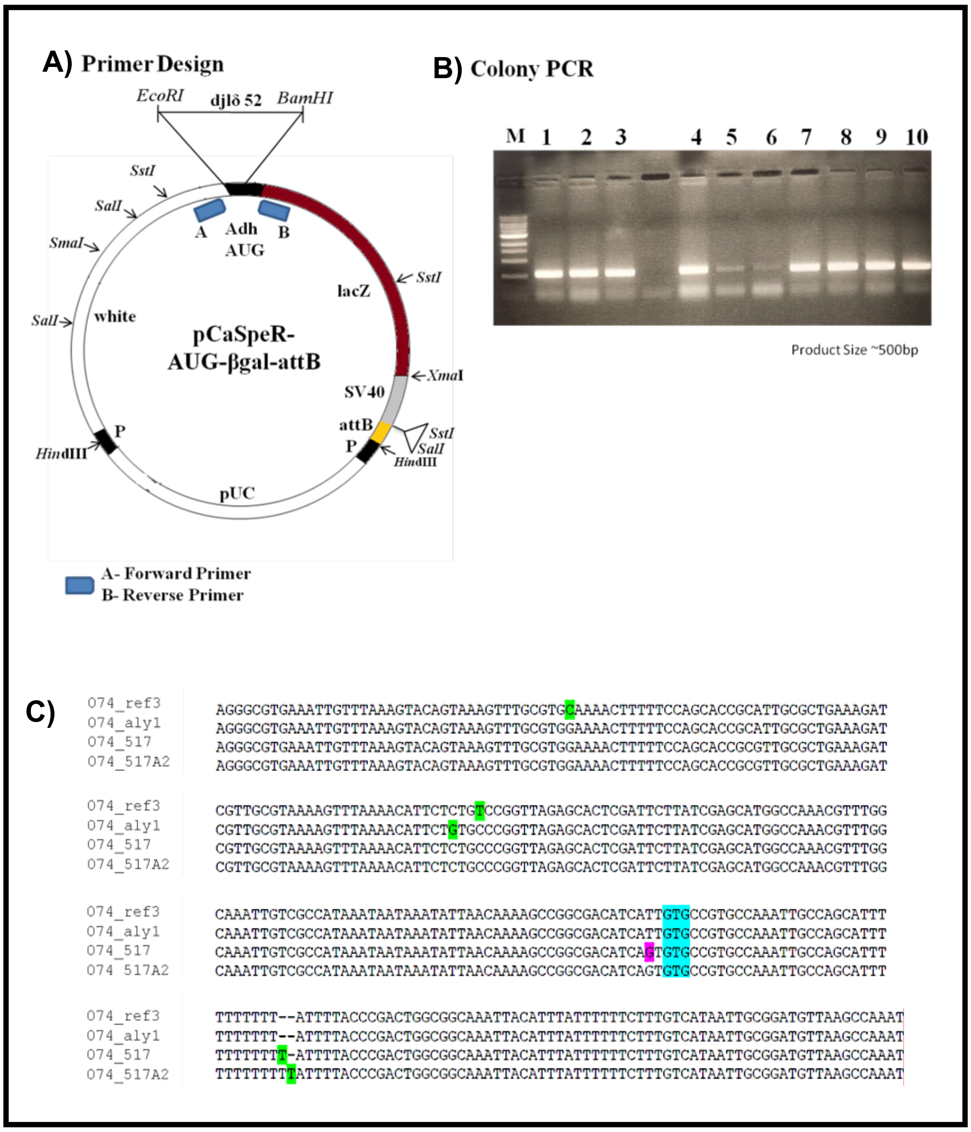


Figure 4.2.2:- A phylogenetic tree for gene_074. A phylogenetic analysis showing expressed alleles are shown by red box and non-expressed alleles are shown by blue box. The promoter is promoter of *aly* gene used as outgroup. The number represents the bootstrap values (%). The tree was generated using maximum likelihood statistical test with 1000 bootstrap iterations. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitution per site. Scale represents distance as percent divergence.

4.3. Cloning of synthetic DNA of gene_074 in reporter vector pCaSpeR-AUG-βGal-attB

Detailed design of synthetic DNA fragments of gene_074 and cloning into pCaSpeR-AUG-βGal-attB is explained in the materials and methods chapter**.** All the synthetic DNA fragments were cloned into reporter vector pCaSpeR-AUG-βGal-attB and a set of primers flanking the multiple cloning site were used to confirm the cloning of insert **(Figure 4.3A)**. Colony PCR was performed using these set of primers and clones were confirmed based on the observed product size (500bp) **(Figure 4.3B)**. Confirmed clones for all the fragments 074_ref3, 517, and aly1 were sequenced by Sanger sequencing (Eurofins Genomics).



```
074_ref3     AGGGCGTGAAATTGTTTAAAGTACAGTAAAGTTTGCGTGCAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGAT
074_aly1     AGGGCGTGAAATTGTTTAAAGTACAGTAAAGTTTGCGTGGAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGAT
074_517      AGGGCGTGAAATTGTTTAAAGTACAGTAAAGTTTGCGTGGAAAACTTTTTCCAGCACCGCGTTGCGCTGAAAGAT
074_517A2    AGGGCGTGAAATTGTTTAAAGTACAGTAAAGTTTGCGTGGAAAACTTTTTCCAGCACCGCGTTGCGCTGAAAGAT

074_ref3     CGTTGCGTAAAAGTTTAAAACATTCTCTGTCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGG
074_aly1     CGTTGCGTAAAAGTTTAAAACATTCTGTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGG
074_517      CGTTGCGTAAAAGTTTAAAACATTCTCTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGG
074_517A2    CGTTGCGTAAAAGTTTAAAACATTCTCTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGG

074_ref3     CAAATTGTCGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCCAGCATTT
074_aly1     CAAATTGTCGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCCAGCATTT
074_517      CAAATTGTCGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCAGTGTGCCGTGCCAAATTGCCAGCATTT
074 517A2    CAAATTGTCGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCAGTGTGCCGTGCCAAATTGCCAGCATTT

074_ref3     TTTTTTT--ATTTTACCCGACTGGCGGCAAATTACATTTATTTTTTCTTTGTCATAATTGCGGATGTTAAGCCAAAT
074_aly1     TTTTTTT--ATTTTACCCGACTGGCGGCAAATTACATTTATTTTTTCTTTGTCATAATTGCGGATGTTAAGCCAAAT
074_517      TTTTTTTT-ATTTTACCCGACTGGCGGCAAATTACATTTATTTTTTCTTTGTCATAATTGCGGATGTTAAGCCAAAT
074_517A2    TTTTTTTTTATTTTACCCGACTGGCGGCAAATTACATTTATTTTTTCTTTGTCATAATTGCGGATGTTAAGCCAAAT
```

Figure 4.3:- Gene_074 synthetic DNA fragments cloning in pCaSpeR-AUG-βGal-attB. A) Reporter vector showing cloning site and primer design used for colony PCR. B) Colony PCR to confirm the cloning of the synthetic DNA fragments in reporter vector (approximate size of products is 500bp each). C) The sequencing results for 074_ref, 517, and aly1. Also, we found that there was a colony for allele 074_517 with extra insertion downstream of the TSS (highlighted in green colour).

The sequences from sequencing results for all the constructs were correct and while analysing the sequences it was found that one clone had a sequence similar to allele 517 but had an extra insertion downstream of TSS and we called that allele 517A2 **(Figure 4.3C)**. Later all these constructs were used for the embryo injection and generation of transgenic flies for these constructs. Testes from these transgenic flies were dissected out and used for beta-galactosidase activity staining.

4.4. Reporter construct comprising of the promoter sequence from non-expressed allele establishes the endogenous expression of gene_074

To understand the molecular mechanism required for allele specific expression of *de novo* gene_074, we examined the promoter regions in a reporter construct. In our lab, it was established that expression of gene_074 does not correlate with SNPs or indels **(Figure 4.2.1)**. Also, phylogenetic analysis of gene_074 showed that this gene could be evolved at least twice. To test the promoter regions of gene_074, we selected the region comprising -200bp upstream of the TSS and +98bp downstream of the TSS (with respect to *D. melanogaster* reference genome sequence) **(Figure 4.4A)**. Initially, one construct was designed 074_ref contains DNA sequence from non-expressing allele (i.e., genomic reference sequence) of gene_074. This experiment allowed me to established that reference genomic sequence is not capable of driving the gene expression in testes. Transgenic flies were generated by injecting vector into embryos of *yw;nos-φ31;attp40* flies and later this vector was inserted at attP40 site via phi-C31 mediated recombination on the second chromosome. Transgenic flies were recovered by selecting for the *w+* marker.

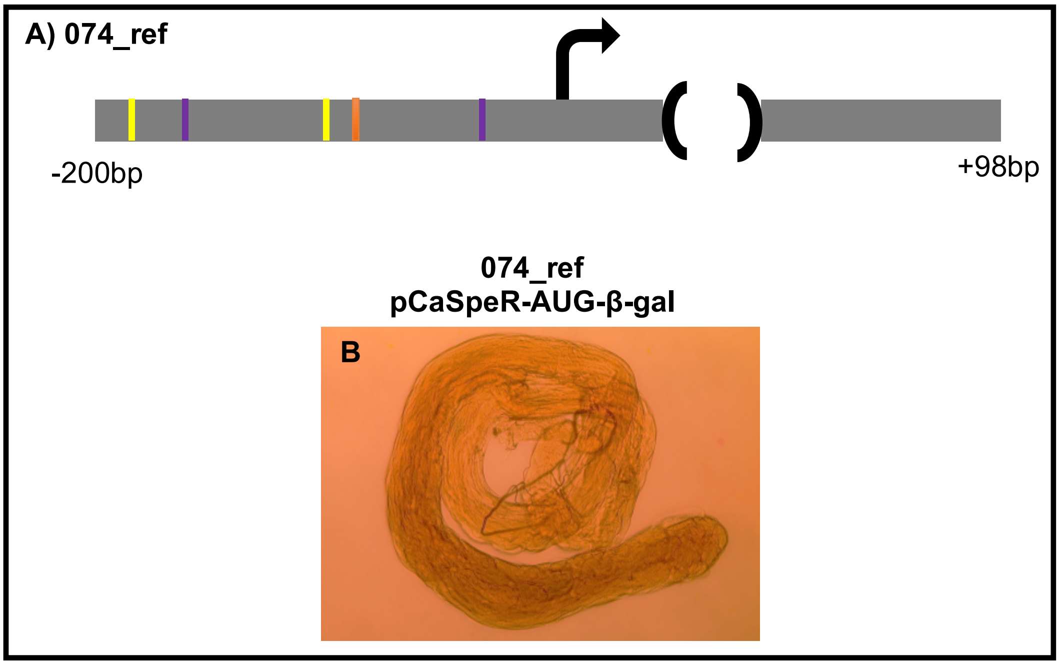


Figure 4.4:- The non-expressing allele of gene_074 gives no expression in testes. A) Construct design for 074_ref showing several SNPs compared to other alleles (yellow-A, purple-T orange-C) upstream of TSS and indel (deletion) downstream of TSS (shown as a bracket). B) Beta-galactosidase activity staining on transgenic testis for this construct show no staining at all. n=pairs of testis.

Beta-galactosidase activity staining on the testes from the transgenic line revealed that construct 074_ref was not capable for driving expression of reporter gene **(Figure 4.4B)**. Now it was established that reference genomic sequence of gene_074 lacks the capability to drive gene expression in testes.

4.5. RAL-517, an expressing allele of gene_074 gives expression in testes

It was established that the non-expressed allele of gene_074 gives no expression in testes. Now, I want to test if expressing allele of gene_074 can give expression in testes. To test this hypothesis we designed a construct 074_517 with sequence with SNPs and indel from expressing line RAL-517 (the only line that expresses gene_074) of gene_074 **(Figure 4.5B)**. Transgenic flies were generated for this transgene and testes were dissected for beta-galactosidase activity staining. Beta-galactosidase activity staining revealed that promoter sequence can drive expression in testes and showed the staining in late spermatids and waste bag of the testes. There was no obvious staining in pre-meiotic stages of spermatogenesis **(Figure 4.5D)**.

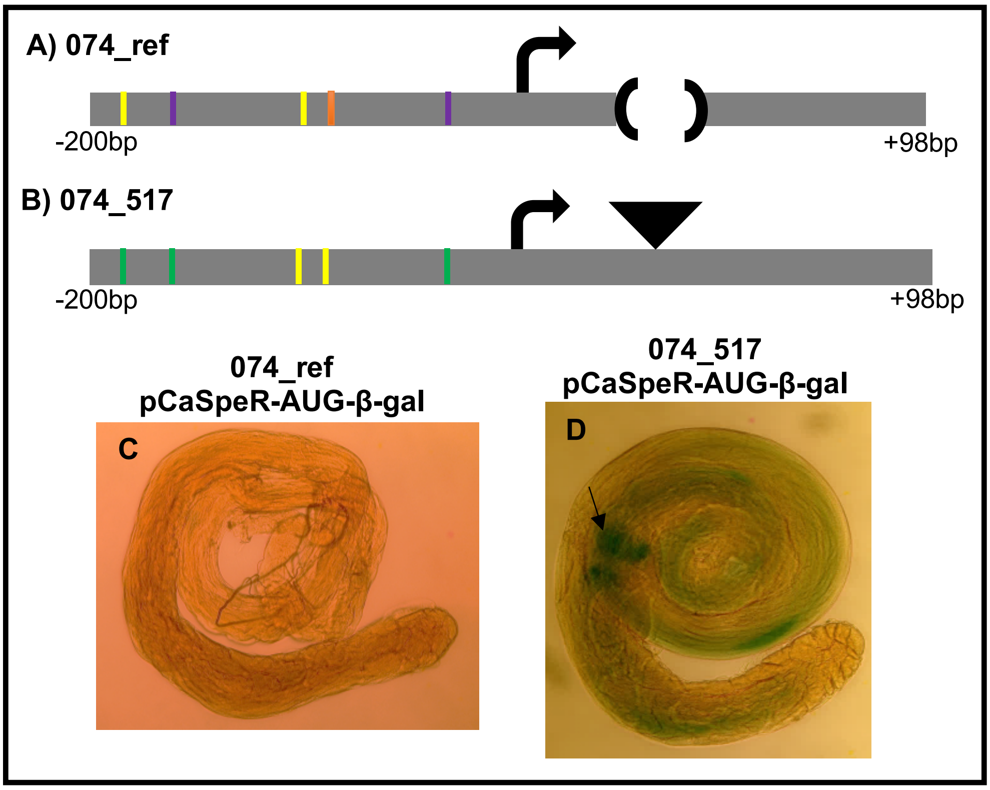


Figure 4.5:- Another expressing allele of gene_074_517 gives expression in testes. A) Design of Construct 074_ref (SNPs are highlighted in colour and indel by a bracket) B) Construct 074_517 comprised of SNPs compared to 074_ref (highlighted in colours yellow-A, orange-C, purple-T green-G) and indel (insertion) shown as a triangle. C) Beta-galactosidase staining on transgenic testes from 074_ref and D) 074_517 construct showed weak staining in early and late spermatids, and waste bag shown by an arrow. n=5 pairs of testis.

4.6. Allele of gene_074 from *aly* chromosome can drive an expression of reporter gene in testes

The allele from *aly* chromosome and non-expressing allele of gene_074 (074_ref) differ by 3 SNPs **(Figure 4.6A&C)** whereas the allele from *aly* chromosome and only expressing allele of gene_074 (517) differ by 3 SNP immediate upstream of the TSS and 1 Indel downstream of the TSS **(Figure 4.6B&C)**. These two expressing alleles and other non-expressing allele differ by 1 SNP and 1 indel. Here, I wanted to test whether the chosen promoter region of a allele of gene_074 from *aly* chromosome is capable of driving the expression of reporter gene in testis. To test this allele, we created one more construct with sequence of allele of gene_074 from *aly* chromosome **(Figure 4.6C)**.

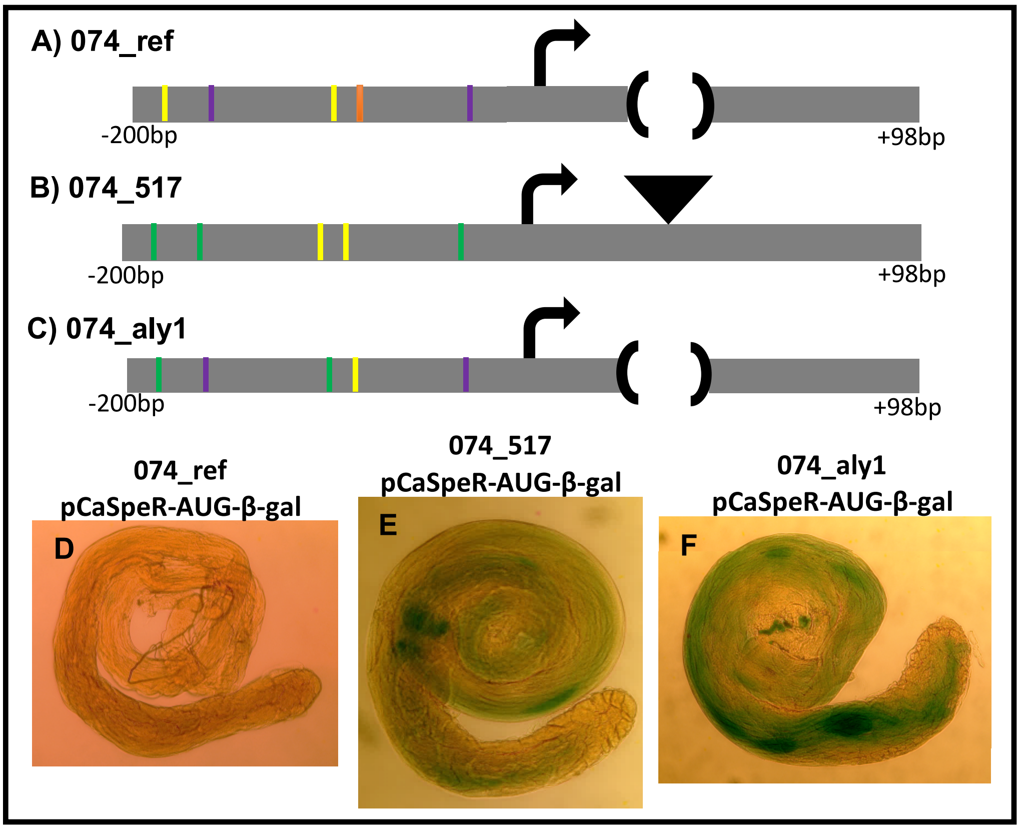


Figure 4.6:- Expressing allele of gene_074 from *aly* chromosome gives expression in testes. A&B) Design of construct 074_ref and construct 074_517. C) Construct 074_aly1 contain SNPs (highlighted in colours yellow-A, purple-T, green-G, and orange-C) and indel (deletion) shown by a bracket. D&E) Beta-galactosidase staining on transgenic testes from 074_ref and 074_517. F) 074_aly1 construct shows strong staining in early and late spermatids but no obvious staining in other cells of spermatogenesis. n=5 pairs of testis.

Testes from this transgene were dissected and subjected to beta-galactosidase activity staining. Beta-galactosidase activity staining discovered that the allele from aly chromosome is capable for driving the expression of the reporter gene in early and late spermatogenesis whereas there was no visible staining observed in germ stem cells (GSCs), spermatogonia, and spermatocytes **(Figure 4.6F)**. Thus, it demonstrates that point mutation (in allele from *aly* chromosome) just upstream of the TSS has no effect on reporter gene expression in testes. Also, the indel i.e., insertion downstream of the TSS (in allele from *aly* chromosome) has no effect on gene expression.

4.7. Additional insertion in allele 517 is also capable of driving the expression in testes

Now it was established that allele 517 and allele from *aly* chromosome of gene_074 are capable of driving the expression of reporter gene in testes. Here I want to test whether the additional insertion caused by sequencing error in allele 517 (later we called this allele as 517A2) of gene_074 is expressed **(Figure 4.3C & 4.7B)**. Testes from transgenic flies for 517A2 construct were dissected and beta-galactosidase activity staining was performed. Beta-galactosidase activity staining showed that an extra insertion downstream of the TSS in allele 517 does not alter the expression of reporter gene and staining was observed in spermatids **(Figure 4.7C)**. There was no visible staining observed in GSCs, spermatogonia, spermatocytes **(Figure 4.7D).**

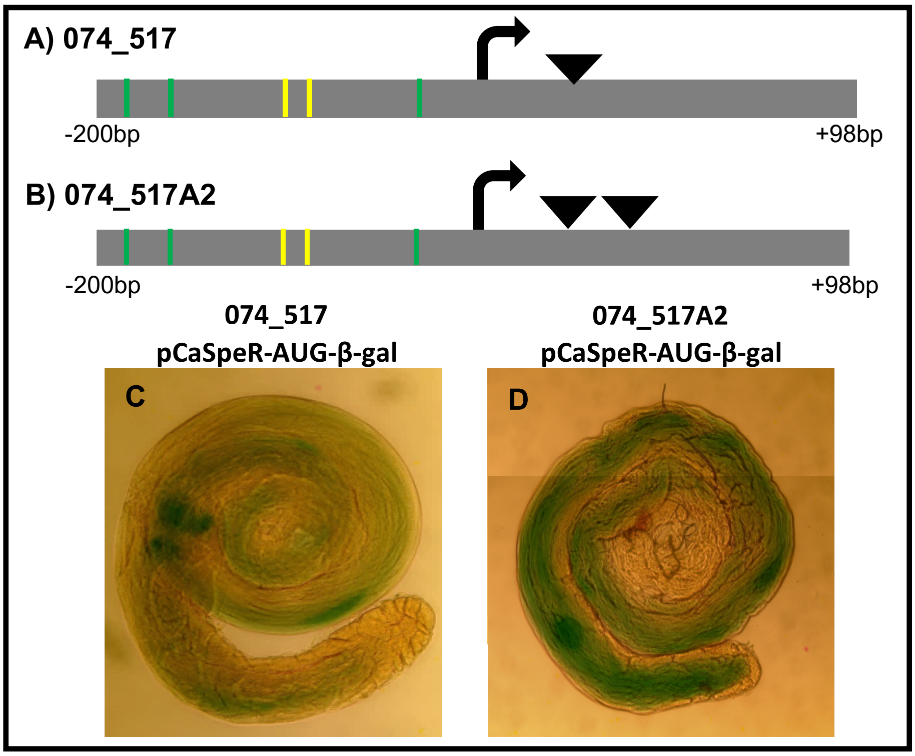


Figure 4.7:- Additional insertion in expressing allele of gene_074_517A2 gives expression in testes. A) Design of construct 074_517 showing SNPs and indel. B) Construct 074_517A2 comprised of SNPs (highlighted in colours yellow-A and green-G) and two indels (insertions) shown as a triangle. C) Beta-galactosidase staining on transgenic testes from 074_517 construct, and D) 074_517A2 construct shows strong staining in spermatids. n=5 pairs of testis.

Altogether these results suggest that tested SNPs or indel are not important for the expression of gene_074 in testes. A combination of different SNPs from expressing and non-expressing alleles could help us understand the mechanism of gene expression regulation.

4.8. Summary

Alignment of the promoter sequences of expressed and non-expressed allele of gene_074 show that no individual SNPs or indels correlate with its expression in testes. Constructs tested in this chapter show that SNPs or indels are not important for expression of gene_074 in testes. An additional insertion in 517 allele has no effect on expression of reporter gene. Two expressing alleles 074_517 and aly1 differ by one SNP which does not affect expression of gene_074 in *Drosophila* testes.

Chapter 5. Results

Recombinant expression and characterisation of subunits of testis-specific meiotic arrest complex (tMAC)

5.1. Abstract

During progression of *Drosophila* spermatogenesis, a large set of genes are expressed in primary spermatocytes regulated by testis-specific meiotic arrest complex (tMAC) along with testis-specific transcription factor II D (tTFIID)(discussed in Introduction section). tMAC is a major transcription complex and yet where it binds in the genome is not well-known (Laktionov et al. 2014). tMAC has four proteins with predicted DNA binding domains and these proteins are Topi, Comr, Tomb, and Achi/Vis. Identifying the binding sites for these proteins and the whole complex in the genome would tell us about how it regulates the gene expression for successful advancement of a vital process of *Drosophila*. In this chapter, I will present the cloning and recombinant protein expression of the full length as well as the DNA binding domain only coding sequences for these four genes. Expressed proteins will be subjected to SELEX for *de novo* discovery of binding site for the respective protein. For each gene we will have two constructs, the first with full length sequence of gene and the second with DNA binding domain sequence. These two different constructs for each gene would help us to find out the binding site of a given protein.

5.2. Topi (matotopetli)

The *topi* gene encodes a 814 amino acid multiple Zn-finger predicted protein with a predicted molecular weight of 92 kDa (Perezgasga et al. 2004) **(Figure 5.1)**. The translated Topi protein contains 10 C2H2 Zn fingers and 1 C2HC finger **(Figure 5.1)** predicted by the PFAM and prosite protein domain analysis tools (El-Gebali et al. 2019, Sigrist et al. 2013). The predicted Zn fingers are clustered in the central region of the protein (amino acids 230-650), with the first two fingers being separated from the remaining nine by a 64 amino acid spacer (Perezgasga et al. 2004). Zn fingers in Topi protein are conserved in *Drosophila melanogaster* orthologue *Drosophila pseudoobscura* (Perezgasga et al. 2004). The last 5 Zn fingers in C-terminal end of Topi protein are more conserved than N-terminal ones and it was the first protein to be cloned and expressed therefore we chose C-terminal end of Topi for SELEX analysis (explained in Chapter 6)

5.2.1. RT-PCR amplification of Topi and sub-cloning to expression vectors

In-order to express C-terminal (C-term), N-terminal (N-term), and full length (FL) of Topi proteins for biochemical characterisation and Protein-DNA interaction analysis by SELEX, 1-2µg/µl of protein must be generated (1-2µg/µl of protein is enough for SELEX). The size of the coding sequence of Topi FL gene is 2445bp, Topi N-terminal is 1473bp, and Topi C-terminal is 969bp. To generate the amplicons for these genes, a pair of testis from wild type flies were dissected and total RNA was extracted. RNA was then converted into cDNA (see Material and Methods) and the nucleotide sequences of the coding regions of *Drosophila* Topi FL, C-term, N-term were amplified from cDNA. For Topi FL, N-term, and C-term the forward primers used were designed to create a NdeI site incorporating a start codon and reverse primers used were designed to create a HindIII site with stop codon **(Figure 5.1)**.

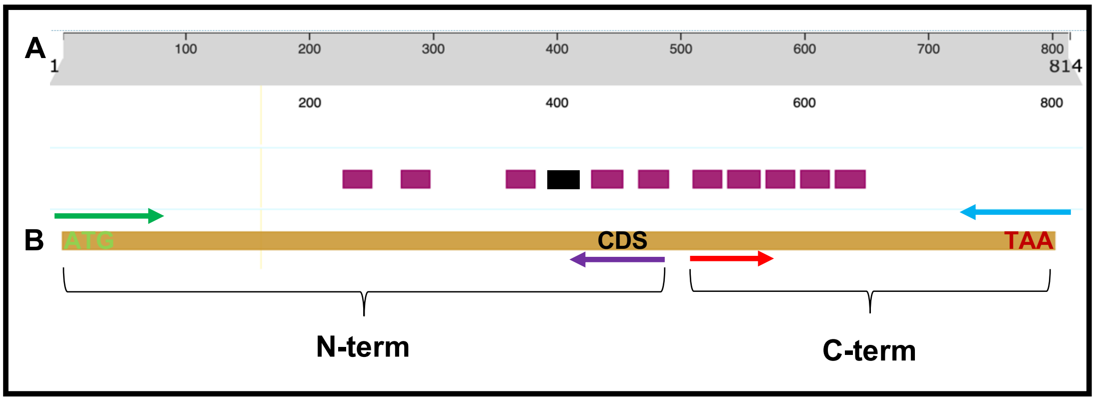


Figure 5.1:- Schematic representation of Zn-Finger domains in Topi and primer design for Topi FL, C-term, and N-term of topi. A) Presentation of Topi FL protein sequence with predicted 11 Zn-finger spread across the protein and 10 C2H2 fingers and 1 C2HC finger are highlighted by purple and black boxes respectively. B) Topi FL CDS showed by orange box and primer set for Topi FL are represented by green and blue arrow. For N-term of topi, sense primer is showed by green and antisense is showed by purple arrow. For C-term of topi, sense primer is showed by red and antisense is showed by blue arrow. Start and stop codon are shown by green and red alphabets.

A table of primers is shown in Appendix, the amplified PCR products were all approximately ~2.5kb, ~1.5kb, and ~970bp, which are in accordance with the length of the topi FL, topi N-term, and topi C-term coding regions respectively. All the PCR products were purified and digested with respective restriction enzymes allowing for cloning into appropriate expression vector pET28a(+) (see Material and methods chapter for detailed cloning procedure). The resulting constructs were transformed into DH5 alpha cells, as described in Material and Methods. A number of colonies were picked for colony PCR and PCR was carried out to check the presence of each gene insert **(Figure 5.2)**. Positive recombinant clones were checked by sequencing.

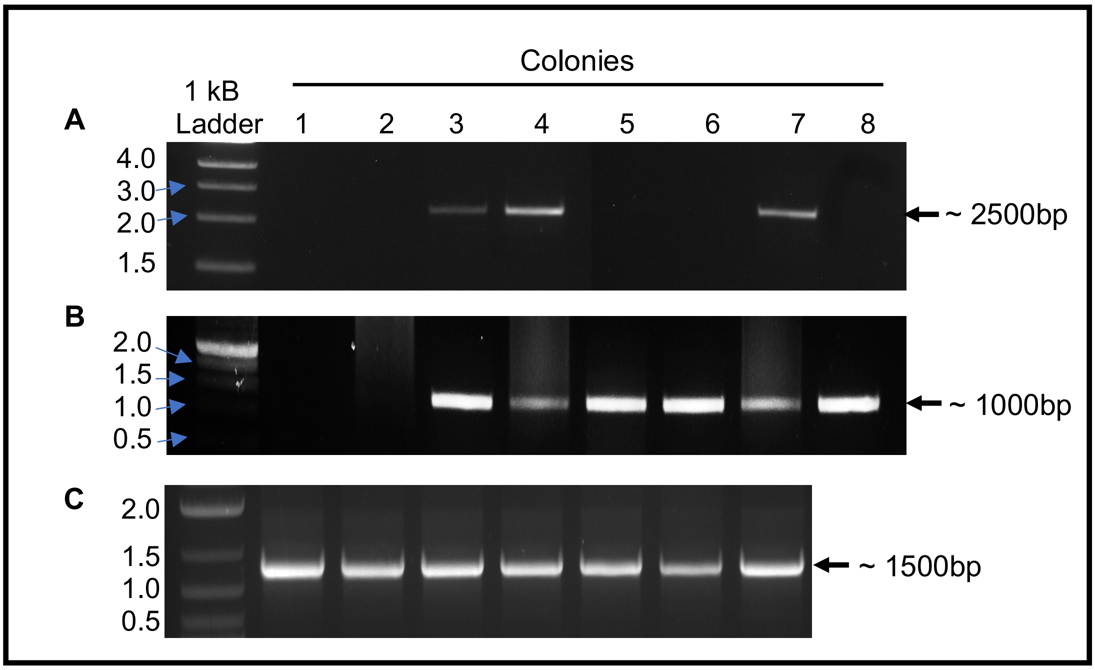
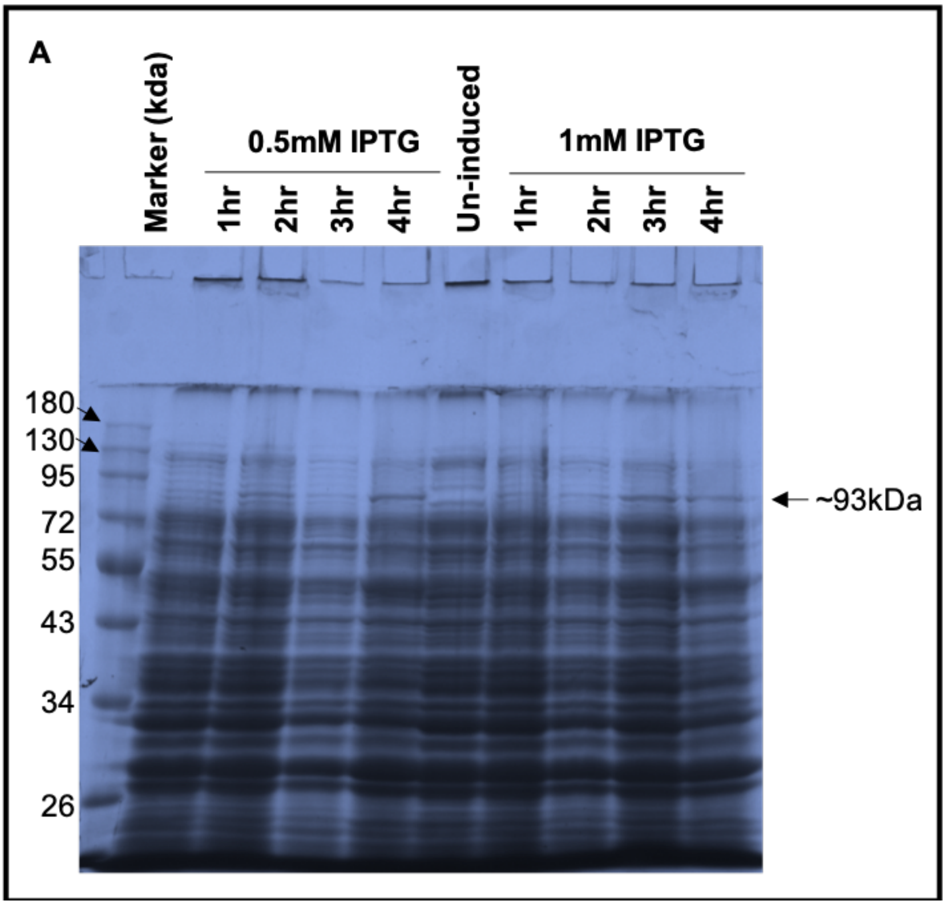
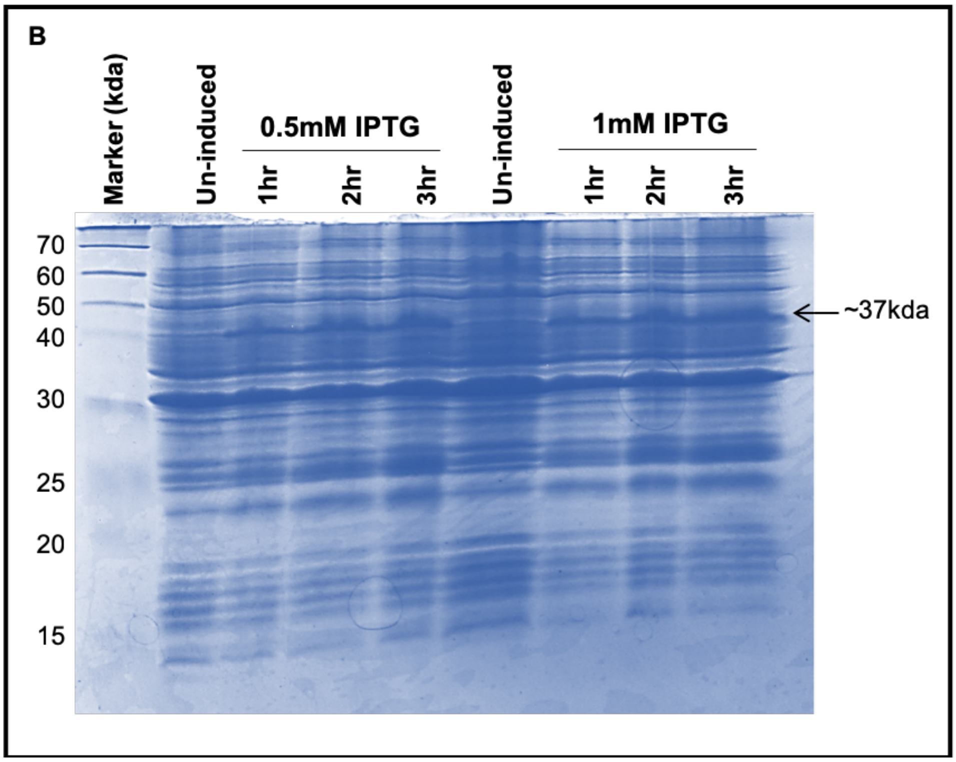


Figure 5.2:- Gel electrophoresis of colony PCR products were amplified using vector specific primers. A) A Panel represents to topi FL colony PCR and three colonies were positive with approximately 2500bp long PCR product. B) A panel representing C-terminal of topi and six colonies were positive with about 970bp PCR product. C) This panel showing colony PCR for N-terminal of topi and seven colonies were positive with about 1500bp long PCR product.

5.2.2. Recombinant over-expression of Topi FL, C-term, and N-term in *E.coli*

The sequence confirmed expression constructs for Topi FL, C-term, and N-term were used to transform the *E. coli* expression host, BL-21(DE3) (see Material and Methods). Initial expression trials for Topi FL, N-term, and C-term were conducted on a small scale (100ml cultures) in order to evaluate the degree of induction and solubility of each protein when expressed as a recombinant protein in *E. coli.* All three recombinant proteins were expressed by addition of 0.5mM and 1mM IPTG and allowed cells to grow for 3hours at 25^0C with rotation with 180rpm. For Topi FL, no protein induction was observed (data not shown). Further to standardise the induction, Topi FL was incubated for 4 hours at 25^0C with rotation of 180rpm (see Material and Methods). After incubation, bacterial cells were harvested and lysed as explained in Material and Methods, and equal concentration of samples were taken from the total lysate and separated and analysed by SDS-PAGE (see Material and Methods). For Topi FL, C-term, and N-term, bands of expected sizes (93, 37, and 56kDa respectively) **(Figure 5.3)** corresponding to $(His)_6$-tagged versions of the proteins of interest, were present in the total lysate of BL-21 cells. For each case, expression construct was induced with 1mM IPTG at 25^{0}C for 3 hours for C-term and N-term and Topi FL was incubated for 4 hours. Cells were harvested and lysed with standard lysis buffer and fractionated by centrifugation (see Material and Methods). Each fraction for the respective protein was separated on SDS-PAGE and enough soluble expression was observed by visualising the enriched band of expected size for respective protein **(Figure 5.4A,B&C)**.





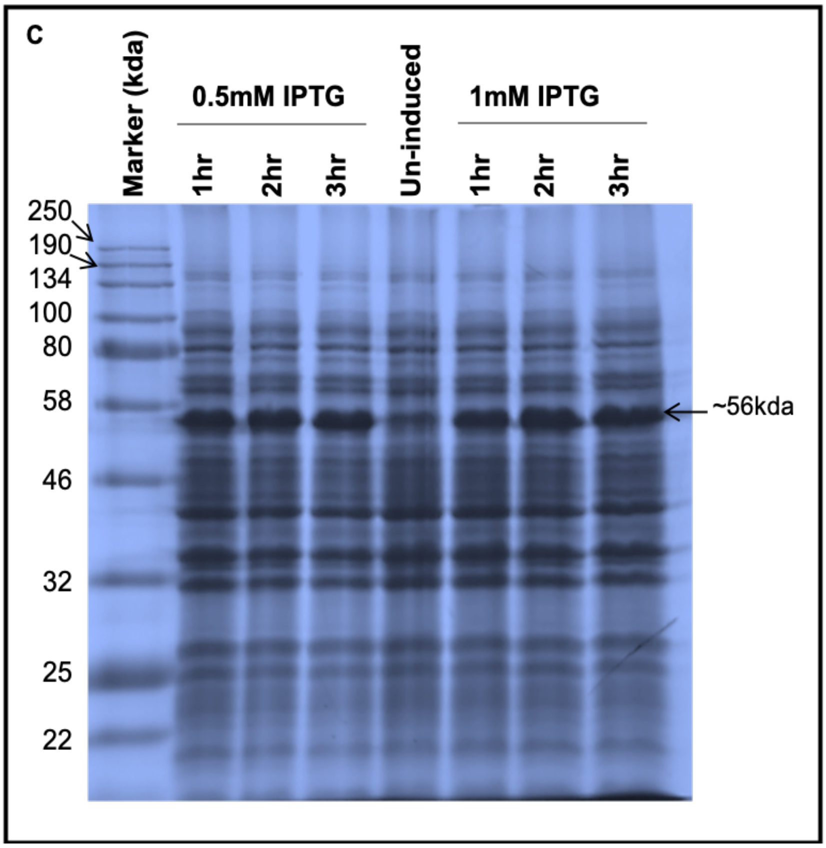


Figure 5.3:- IPTG induced over expression of His-tagged Topi FL, C-term, and N-term in *E.coli.* Coomassie stained gels showing A) Topi FL induction at 0, 1, 2, 3, and 4hrs using 0.5mM and 1mM IPTG. ~93kDa band was observed in gel (Marker NEB P). B) Topi C-term over-expression was carried out at 0, 1, 2, 3hrs using 0.5mM and 1mM IPTG and band of 37kDA was observed. C) N-term of Topi over-expression was carried out at 0, 1, 2, 3hrs using 0.5mM and 1mM IPTG and the band of 56kDa was observed in gel.

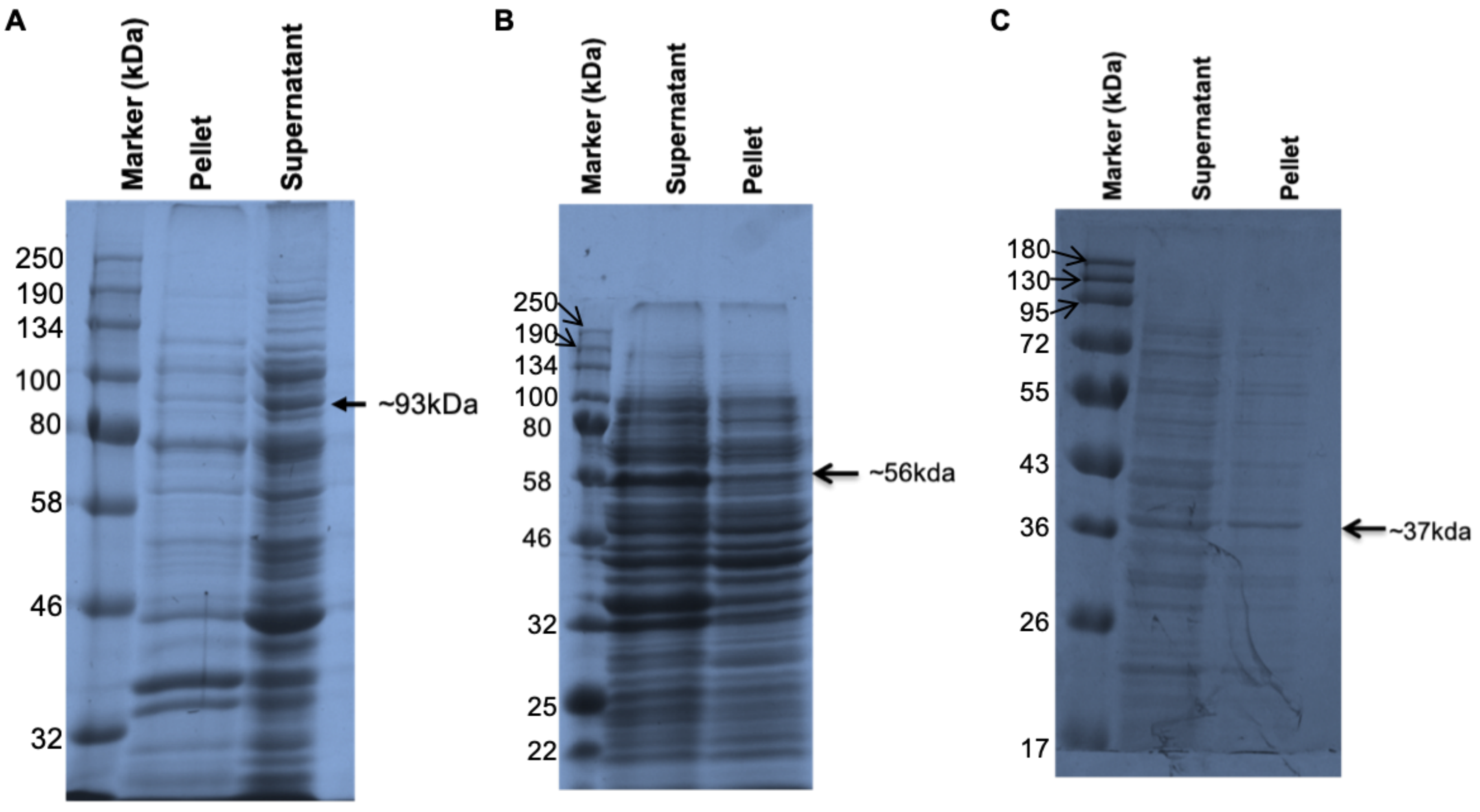


Figure 5.4:- Soluble fractionation of protein His-tagged Topi FL, N-term, and C-term. A) Fractionation of protein Topi FL showing that soluble fraction has enough protein of interest (band of 93kDa) and very little was observed in pellet fraction. B) For N-term of Topi, 56kDa of band was observed in soluble fraction and C) for C-term 37kDa band was observed in soluble fraction (Markers NEB P7717, P7718, P7719).

5.3. Comr (Cookie monster)

comr gene encodes a 600 amino acids Winged helix DNA-binding domain superfamily protein with a molecular weight of 68.4 kDa (Jiang and White-Cooper 2003) **(Figure 5.5)**. PFAM and prosite domain analysis tools predicted Comr protein contains HSF type DNA binding domain (amino acids 113-209) **(Figure 5.5)**.

5.3.1. Cloning of Comr FL and DBD, and recombinant over-expression of each protein in *E. coli* expression host

For generation of PCR products, nucleotide sequences for Comr FL and Comr DBD was amplified from cDNA using suitable primer set **(Figure 5.5)** (see primer list). The expected PCR product for Comr FL is 1800bp long and Comr DBD is 360bp. The PCR products generated for respective genes were purified and digested with respective restriction enzymes and cloned into expression vector pET28a(+). Cloned inserts were confirmed by selecting number of colonies for each gene and performed colony PCR using T7 primer set (see primer list) **(Figure 5.6)** followed by sequencing.

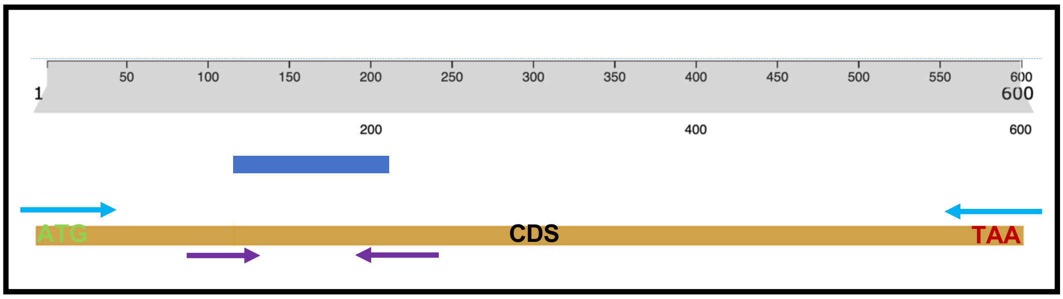


Figure 5.5:- Schematic representation of Comr full length protein as well as DNA binding domain and primer design for Comr FL and Comr DBD. Comr FL protein is made up of 600 amino acids and DBD comprises amino acid sequence from 113-209 (shown by blue box). Primer set for Comr FL are shown by blue arrows and primer set for Comr DBD are shown by purple colour. Start and stop codon are shown by green and red letters.

The sequence confirmed expression vectors for Comr FL and DBD were transformed into *E. coli* expression host BL-21(DE3). For protein induction, single colony for each vector was incubated in 100ml of LB plus Kanamycin and protein expression was induced by addition of 0.5mM and 1mM IPTG and cells were allowed to grow for 3hrs at 25^0C with rotation at 250rpm. Cells were harvested by centrifugation at 5000rpm and from whole cell lysate, equal concentration of proteins were used for SDS-PAGE and protein expression was confirmed **(Figure5.7)**. The expected size of bands for full length Comr protein (70kDa) and Comr DBD (12kDa) were observed in both induced conditions

(Figure 5.7A&B). In-order to check the solubility of each protein, each expression vector was induced using 1mM IPTG and incubated at 25^{0}C for 3hrs with rotation at 180rpm and harvested cells were lysed using the standard lysis buffer (see Material and Methods). An equal concentration of protein from supernatant and pellet fraction was taken and solubility of each protein was confirmed by SDS-PAGE analysis **(Figure 5.8A&B)**.

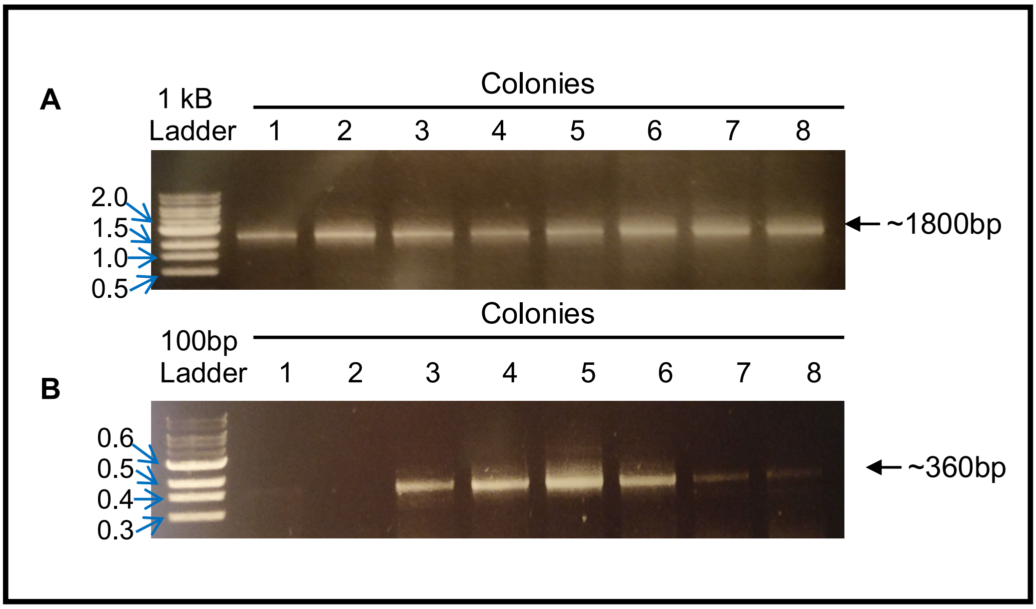
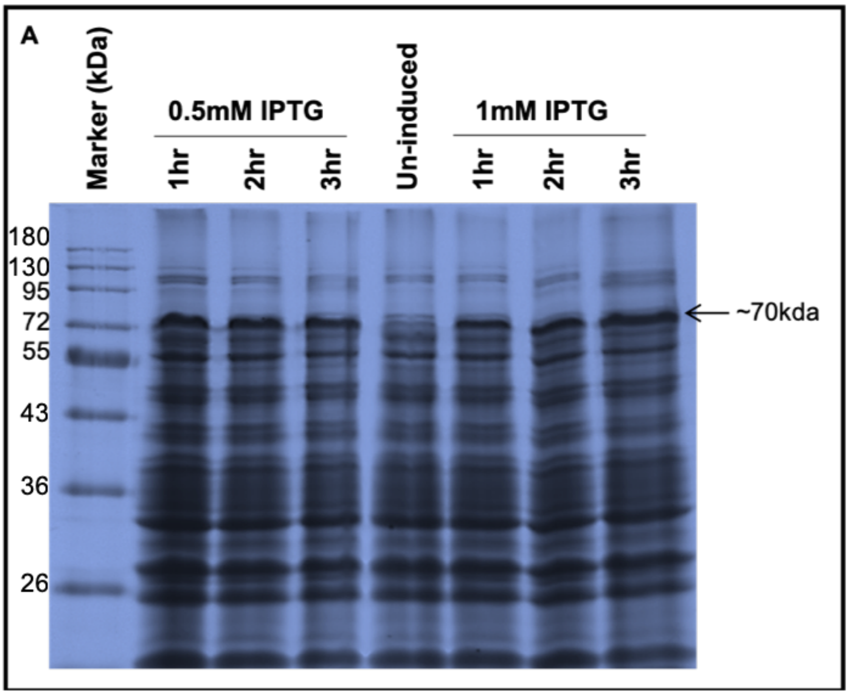


Figure 5.6 Gel electrophoresis of colony PCR products were amplified using vector specific primers(T7 promoter) for Comr FL and Comr DBD. A) A panel represents to Comr FL colony PCR and eight colonies were positive with approximately ~1800bp long PCR product. Bottom panel showing colony PCR for Comr DBD and seven colonies were positive with about ~360bp long PCR product.



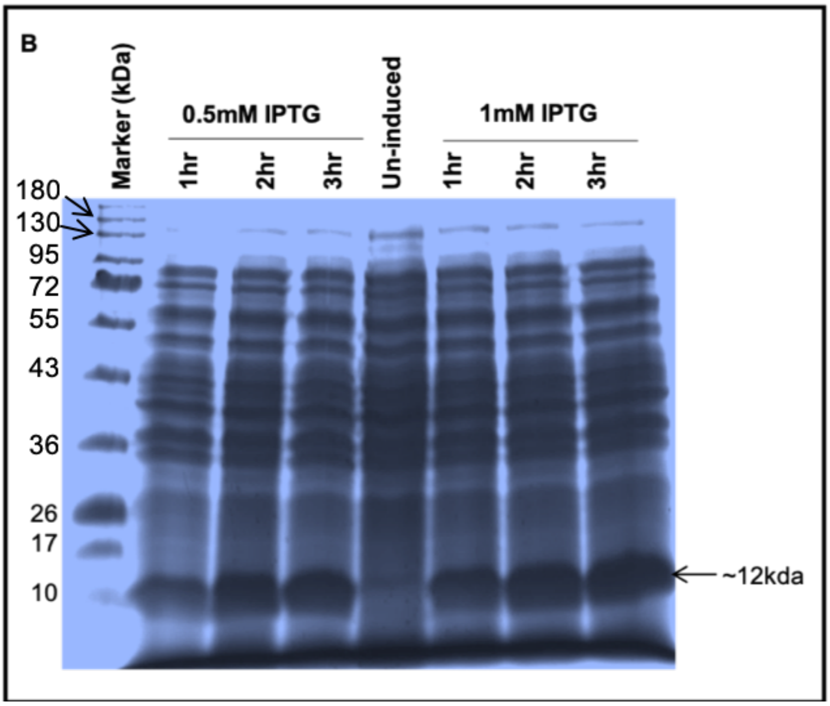


Figure 5.7:- IPTG induced over-expression analysis of His-tagged proteins Comr FL and Comr DBD. Coomassie stained gels showing A) Over-expression was carried out for 0 (uninduced), 1, 2, 3hrs using 0.5mM and 1mM IPTG and band of ~70kDa was observed (Marker NEB P7718). B) Comr DBD carried out for 0, 1, 2, 3hrs using 0.5mM and 1mM IPTG and band of ~12kDa was observed for Comr DBD (Marker NEB P7719).

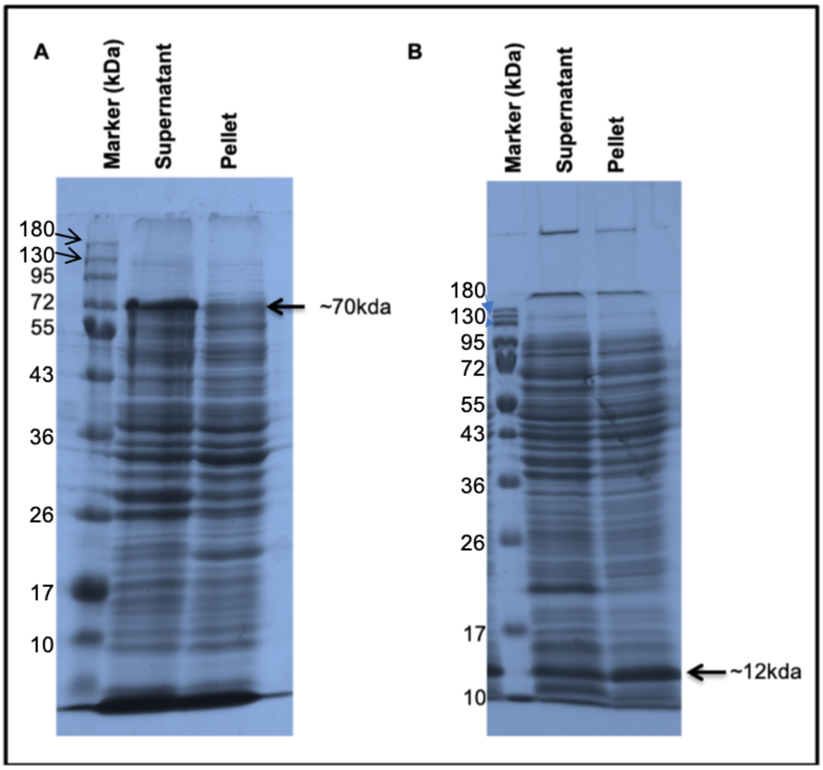


Figure 5.8:- Fractionation of Comr FL and Comr DBD proteins. A) Soluble fraction had enough amount of Comr FL compare to pellet and band of 70kDa was observed (Marker NEB P7718). B) Soluble fraction of Comr DBD had enough protein and band of about 12kDa was observed (Marker NEB P7719).

5.4. Tomb (Tombola)

The *tomb* gene encodes a 243 amino acids tesmin/TSO1 family protein with a molecular weight of 26.18kDa (Jiang et al. 2007). Tomb protein contains a CXC DNA binding domain (amino acids 1-63) predicted by prosite and PFAM domain analysis tools **(Figure 5.9)**.

5.4.1. Tomb FL and DBD cloning and protein expression in *E. coli*

cDNA was amplified using appropriate primer sets for Tomb FL and Tomb DBD and the resulting PCR product was purified. Purified PCR products for each gene were digested with respective restriction enzymes and cloned into the expression vector pET28a(+). To confirm the cloning, a number of colonies for each product were picked and colony PCR was performed using T7 primer set and positive colonies were sequenced and inserts were confirmed **(Figure 5.10)**.

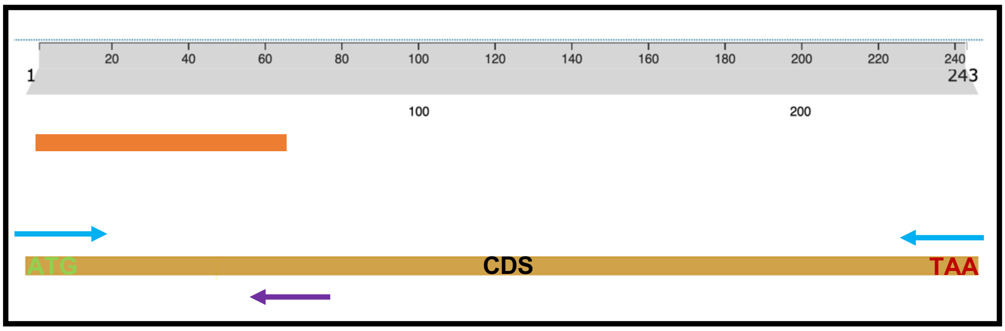


Figure 5.9:- Representation of Tomb protein sequence and primer design. Tomb FL protein is 243 amino acids long and has DBD (orange block) domain comprises sequence from 1-63 amino acids. Primers for full length protein were designed and shown by blue arrows and primers for DBD are shown by forward primer as a blue arrow and reverse primer as a purple arrow.

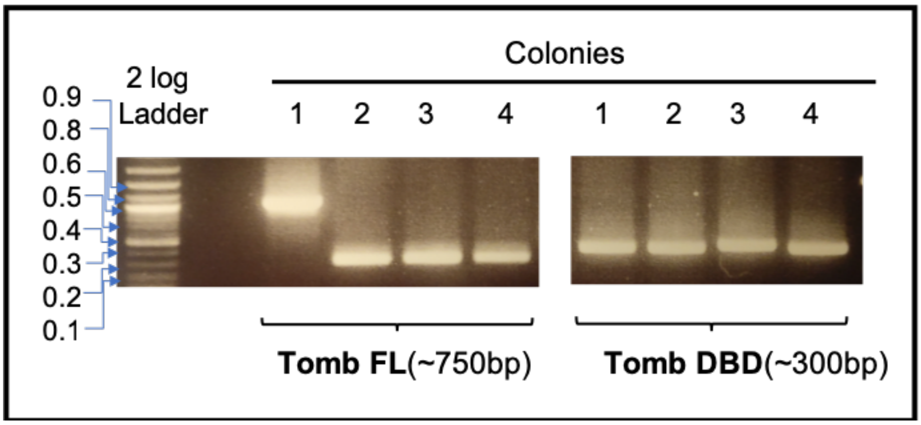


Figure 5.10 Colony PCR products were amplified using vector specific primers (T7 promoter) for Tomb FL and Tomb DBD. Left hand panel represents to Tomb FL colony PCR and one colony was positive with approximately ~750bp long PCR product. Right hand panel showing colony PCR for Tomb DBD and four colonies were positive with about ~300bp long PCR product.

The sequence confirmed expression vector for each gene were selected for protein induction and transformed in BL-21(DE3) and incubated overnight at 37^{0}C. For Tomb FL and Tomb DBD protein induction, a single colony for each region was incubated in 100ml of LB plus *Kanamycin* and protein expression was induced by addition of 0.5mM and 1mM of IPTG. Cells were allowed to grow for 3hrs at 25^{0}C with rotation at 180rpm and after incubation cells were harvest by centrifugation at 5000 rpm. Whole cell lysate was used for analysis by taking equal concentration for each protein and protein induction was confirmed by observing the expected band size of His-tagged version of protein Tomb FL and Tomb DBD (11kDA) in SDS-PAGE analysis **(Figure 5.11 & 5.12A)**. For Tomb FL, no protein induction was observed **(Figure 5.11)**. To test the solubility of Tomb DBD, Tomb DBD expression vector was induced using 1mM IPTG for 3hrs at 25^{0}C with rotation at 180rpm and cells were harvested and lysed with standard buffer (see Materials and Methods). Equal concentration of protein was taken from samples (supernatant and pellet) and separated by SDS-PAGE and the significant amount of His-tagged version of Tomb DBD protein was observed with expected band size ~11kDa **(Figure 5.12B)**.

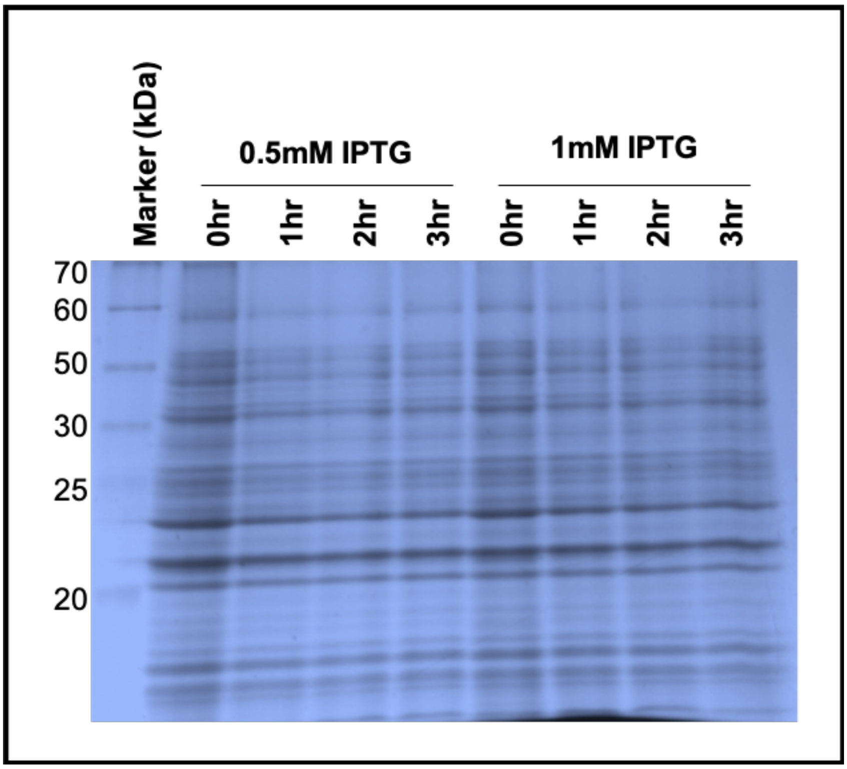


Figure 5.11:- His-tagged Tomb FL protein expression induction. 0.5mM or 1mM IPTG was used to induced the expression of Tomb FL protein for 3hrs and 0hr was uninduced cell lysate. There was no visible over-expression band was observed in both conditions 0.5mM or 1mM IPTG. Expected size of band was ~27kDa.

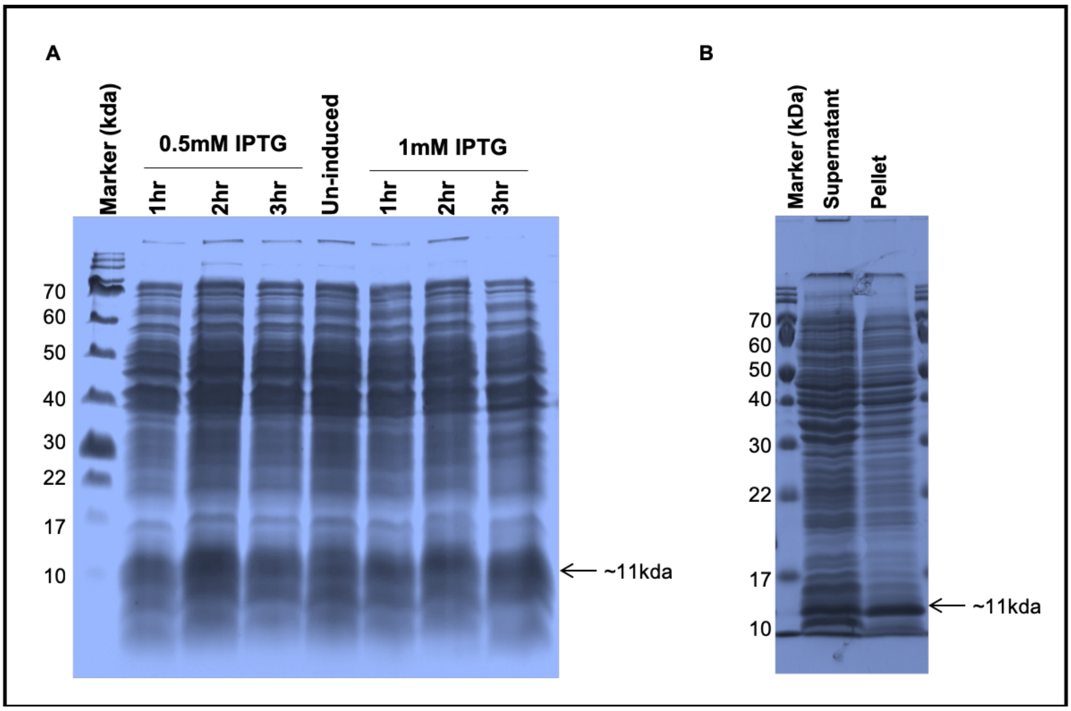


Figure 5.12:- His-tagged Tomb DBD induction and fractionation. A) Coomassie stained gel showing Tomb DBD induction using 0.5mM or 1mM IPTG for 0 (uninduced), 1, 2 ,3hr and expected band of ~11kDa was observed. B) Upon fractionation of his-tagged Tomb DBD protein, expect band of 11kDa was observed in supernatant as well as in pellet.

5.5. Achi/Vis (Achintya/Vismay)

The *achi* gene encodes three different transcripts and these transcripts translate into three proteins, one is 426 amino acids protein (achi-PA) and other two variants (achi-PC & achi-PD) encode 555 amino acids proteins. The size difference between these protein variants is because of an additional protein coding exon in achi-PC and achi-PD. Variant achi-PA and achi-PC are highly expressed in *Drosophila* testes, hence using the full length version which include this additional exon for our experiments. *vis* gene encode two different 424 (vis-PA) and 524 (vis-PB) amino acids proteins. Like *achi*, the difference in size of these two variants is due to an additional protein coding exon in vis-PB, thus using the full length version which includes this exon for experiments. The two genes are highly similar (93% at nucleotide level and 97% at protein level) (Ayyar et al. 2003).

Homeo domain in Achi protein spans amino acids 91-154 **(Figure 5.13)** and in Vis protein, homeo domain spans 89-152 **(Figure 5.14)**. Homeo domain binds DNA through a helix-turn-helix (HTH) structure of protein **(Figure 5.15)**.

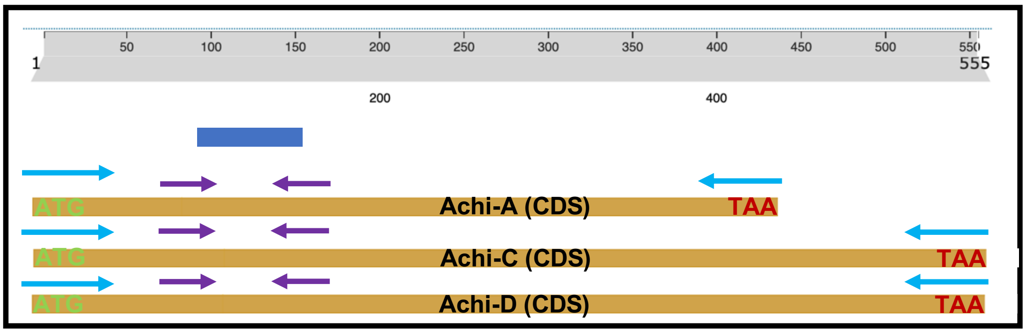


Figure 5.13:- Schematic representation of Achi protein sequence, DBD, and set of primers. The Achi-A is 426 amino acids protein and Achi-C&D are 555 amino acids proteins. The DBD in all the three variants comprise the sequence from 91-154 amino acids. Primers for full length protein are shown by blue colour whereas for DBD are shown by purple colour.

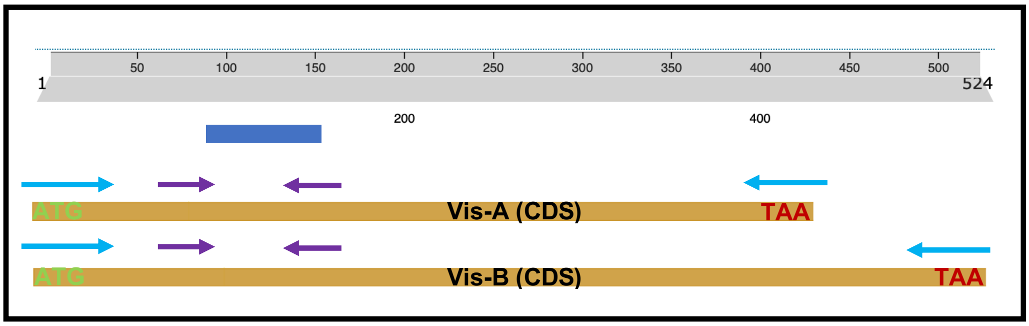
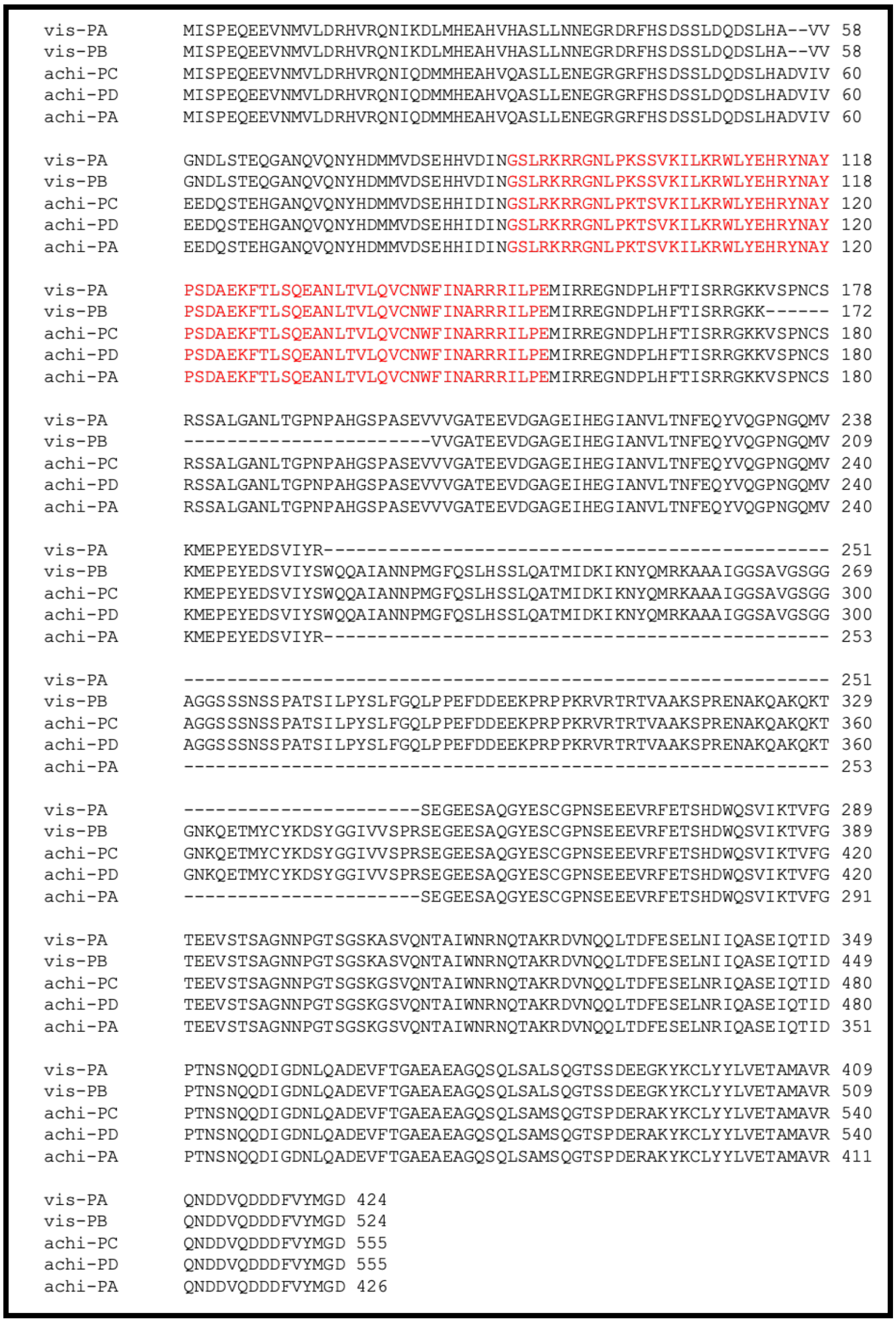


Figure 5.14:- Schematic representation of Vis protein sequence, DBD, and set of primers. The Vis-A is 424 amino acids protein and Vis-B is 524 amino acids protein. The DBD in both variants comprise sequence from 89-152 amino acids. Primers for full length protein are shown by blue colour whereas for DBD are shown by purple colour.

5.5.1. Vis B, Achi DBD, and Vis DBD cloning and over-expression of proteins in *E. coli*

For cloning, cDNA was amplified using appropriate primer sets for Vis B, Achi DBD, and Vis DBD and PCR products were purified. Nucleotide sequence length for Vis B is 1575bp, Achi and Vis DBD 470bp each. Purified PCR products for each gene were digested with respective restriction enzymes and cloned into restriction digested expression vector pET28a(+) (see Material and Methods). Cloned vectors transformed into DH5aplha cells and allowed to grow overnight at 37^{0}C. A number of colonies were selected and colony PCR performed using T7 primer set and the expected band size were observed **(Figure 5.16)**. The clones were confirmed by sequencing and the sequence confirmed expression vectors were selected for protein expression analysis. For expression of proteins, expression vectors for Vis B, Achi DBD, and Vis DBD were transformed in BL-21(DE3) and allowed to grow on Kanamycin agar plates overnight at

37^{0}C(see Material and Methods). Single colony for each expression vector was picked and incubated in 100ml of LB plus Kanamycin and protein expression was induced by addition of 0.5mM or 1mM IPTG and cells were allowed to grow for 3hrs at 25^{0}C with rotation at 180rpm. After incubation, cells were harvested and total lysate was used for protein expression analysis. An equal concentrations of each sample was taken and separated by SDS-PAGE and expected bands of were observed for over-expression of Vis B (60kDa), Achi DBD (14kDa), and Vis DBD (1kDa) **(Figure 5.17A, 5.18A, 5.19A)**.



```
vis-PA       MISPEQEEVNMVLDRHVRQNIKDLMHEAHVHASLLNNEGRDRFHSDSSLDQDSLHA--VV 58
vis-PB       MISPEQEEVNMVLDRHVRQNIKDLMHEAHVHASLLNNEGRDRFHSDSSLDQDSLHA--VV 58
achi-PC      MISPEQEEVNMVLDRHVRQNIQDMMHEAHVQASLLENEGRGRFHSDSSLDQDSLHADVIV 60
achi-PD      MISPEQEEVNMVLDRHVRQNIQDMMHEAHVQASLLENEGRGRFHSDSSLDQDSLHADVIV 60
achi-PA      MISPEQEEVNMVLDRHVRQNIQDMMHEAHVQASLLENEGRGRFHSDSSLDQDSLHADVIV 60

vis-PA       GNDLSTEQGANQVQNYHDMMVDSEHHVDINGSLRKRRGNLPKSSVKILKRWLYEHRYNAY 118
vis-PB       GNDLSTEQGANQVQNYHDMMVDSEHHVDINGSLRKRRGNLPKSSVKILKRWLYEHRYNAY 118
achi-PC      EEDQSTEHGANQVQNYHDMMVDSEHHIDINGSLRKRRGNLPKTSVKILKRWLYEHRYNAY 120
achi-PD      EEDQSTEHGANQVQNYHDMMVDSEHHIDINGSLRKRRGNLPKTSVKILKRWLYEHRYNAY 120
achi-PA      EEDQSTEHGANQVQNYHDMMVDSEHHIDINGSLRKRRGNLPKTSVKILKRWLYEHRYNAY 120

vis-PA       PSDAEKFTLSQEANLTVLQVCNWFINARRRILPEMIRREGNDPLHFTISRRGKKVSPNCS 178
vis-PB       PSDAEKFTLSQEANLTVLQVCNWFINARRRILPEMIRREGNDPLHFTISRRGKK------ 172
achi-PC      PSDAEKFTLSQEANLTVLQVCNWFINARRRILPEMIRREGNDPLHFTISRRGKKVSPNCS 180
achi-PD      PSDAEKFTLSQEANLTVLQVCNWFINARRRILPEMIRREGNDPLHFTISRRGKKVSPNCS 180
achi-PA      PSDAEKFTLSQEANLTVLQVCNWFINARRRILPEMIRREGNDPLHFTISRRGKKVSPNCS 180

vis-PA       RSSALGANLTGPNPAHGSPASEVVVGATEEVDGAGEIHEGIANVLTNFEQYVQGPNGQMV 238
vis-PB       -----------------------VVGATEEVDGAGEIHEGIANVLTNFEQYVQGPNGQMV 209
achi-PC      RSSALGANLTGPNPAHGSPASEVVVGATEEVDGAGEIHEGIANVLTNFEQYVQGPNGQMV 240
achi-PD      RSSALGANLTGPNPAHGSPASEVVVGATEEVDGAGEIHEGIANVLTNFEQYVQGPNGQMV 240
achi-PA      RSSALGANLTGPNPAHGSPASEVVVGATEEVDGAGEIHEGIANVLTNFEQYVQGPNGQMV 240

vis-PA       KMEPEYEDSVIYR----------------------------------------------- 251
vis-PB       KMEPEYEDSVIYSWQQAIANNPMGFQSLHSSLQATMIDKIKNYQMRKAAAIGGSAVGSGG 269
achi-PC      KMEPEYEDSVIYSWQQAIANNPMGFQSLHSSLQATMIDKIKNYQMRKAAAIGGSAVGSGG 300
achi-PD      KMEPEYEDSVIYSWQQAIANNPMGFQSLHSSLQATMIDKIKNYQMRKAAAIGGSAVGSGG 300
achi-PA      KMEPEYEDSVIYR----------------------------------------------- 253

vis-PA       ------------------------------------------------------------ 251
vis-PB       AGGSSSNSSPATSILPYSLFGQLPPEFDDEEKPRPPKRVRTRTVAAKSPRENAKQAKQKT 329
achi-PC      AGGSSSNSSPATSILPYSLFGQLPPEFDDEEKPRPPKRVRTRTVAAKSPRENAKQAKQKT 360
achi-PD      AGGSSSNSSPATSILPYSLFGQLPPEFDDEEKPRPPKRVRTRTVAAKSPRENAKQAKQKT 360
achi-PA      ------------------------------------------------------------ 253

vis-PA       ----------------------SEGEESAQGYESCGPNSEEEVRFETSHDWQSVIKTVFG 289
vis-PB       GNKQETMYCYKDSYGGIVVSPRSEGEESAQGYESCGPNSEEEVRFETSHDWQSVIKTVFG 389
achi-PC      GNKQETMYCYKDSYGGIVVSPRSEGEESAQGYESCGPNSEEEVRFETSHDWQSVIKTVFG 420
achi-PD      GNKQETMYCYKDSYGGIVVSPRSEGEESAQGYESCGPNSEEEVRFETSHDWQSVIKTVFG 420
achi-PA      ----------------------SEGEESAQGYESCGPNSEEEVRFETSHDWQSVIKTVFG 291

vis-PA       TEEVSTSAGNNPGTSGSKASVQNTAIWNRNQTAKRDVNQQLTDFESELNIIQASEIQTID 349
vis-PB       TEEVSTSAGNNPGTSGSKASVQNTAIWNRNQTAKRDVNQQLTDFESELNIIQASEIQTID 449
achi-PC      TEEVSTSAGNNPGTSGSKGSVQNTAIWNRNQTAKRDVNQQLTDFESELNRIQASEIQTID 480
achi-PD      TEEVSTSAGNNPGTSGSKGSVQNTAIWNRNQTAKRDVNQQLTDFESELNRIQASEIQTID 480
achi-PA      TEEVSTSAGNNPGTSGSKGSVQNTAIWNRNQTAKRDVNQQLTDFESELNRIQASEIQTID 351

vis-PA       PTNSNQQDIGDNLQADEVFTGAEAEAGQSQLSALSQGTSSDEEGKYKCLYYLVETAMAVR 409
vis-PB       PTNSNQQDIGDNLQADEVFTGAEAEAGQSQLSALSQGTSSDEEGKYKCLYYLVETAMAVR 509
achi-PC      PTNSNQQDIGDNLQADEVFTGAEAEAGQSQLSAMSQGTSPDERAKYKCLYYLVETAMAVR 540
achi-PD      PTNSNQQDIGDNLQADEVFTGAEAEAGQSQLSAMSQGTSPDERAKYKCLYYLVETAMAVR 540
achi-PA      PTNSNQQDIGDNLQADEVFTGAEAEAGQSQLSAMSQGTSPDERAKYKCLYYLVETAMAVR 411

vis-PA       QNDDVQDDDFVYMGD 424
vis-PB       QNDDVQDDDFVYMGD 524
achi-PC      QNDDVQDDDFVYMGD 555
achi-PD      QNDDVQDDDFVYMGD 555
achi-PA      QNDDVQDDDFVYMGD 426
```

Figure 5.15:- Protein sequence alignment between Achi and Vis showing the homeo domain. Homeo domain is highlighted in red alphabet. Homeo domain in all three variants of achi spans from 91-154 and in both the variants of vis spans from 89-152 amino acids.

To test the solubility of proteins, expression vectors for Vis B, Achi DBD, and Vis DBD were induced using 1mM IPTG at 25^{0}C for 3hrs with rotation of 180rpm. After incubation, cells were harvested and lysed with standard buffer and fractionated by centrifugation (see Material and Methods). An equal concentration of each protein were taken from samples and separated by SDS-PAGE. The enough amount of protein was enriched in supernatant fraction for Vis B **(Figure 5.17B)** and for Achi DBD **(Figure 5.18B)** and Vis DBD **(Figure 5.19B)** proteins were enriched in both fractions, supernatant and pellet.

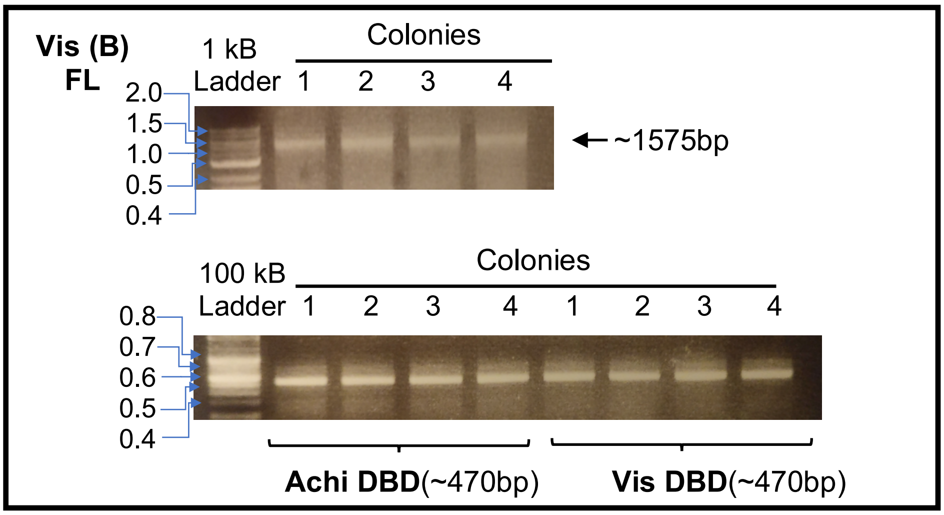


Figure 5.16:- Colony PCR products were amplified using vector specific primers (T7 promoter) for Vis(B) FL, Achi DBD, and Vis DBD. Top panel represents to Vis(B) FL colony PCR and four colonies were positive with approximately ~1575bp long PCR product. Bottom panel showing colony PCR for Achi DBD and Vis DBD and four colonies were positive with about ~470bp long PCR product for each construct.

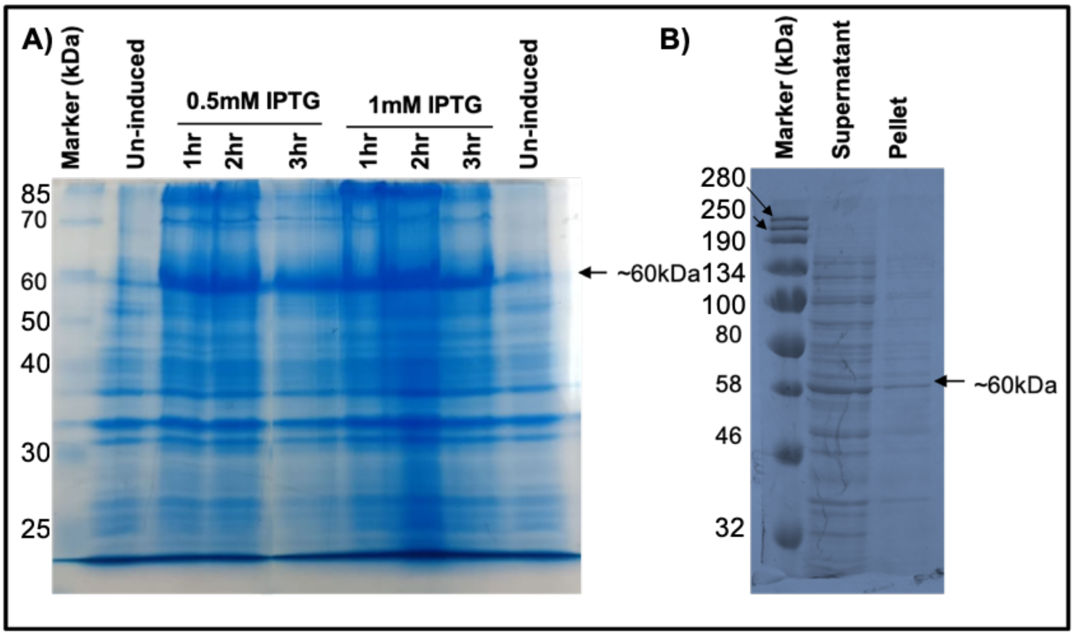


Figure 5.17:- Vis (B) FL induction and fractionation. A) Vis (B) FL protein induction using 0.5mM or 1mM IPTG for 3hrs and expected band of size ~60kDa was observed in both induced conditions. B) Fractionation of Vis (B) FL shows significant amount of protein in soluble fraction.

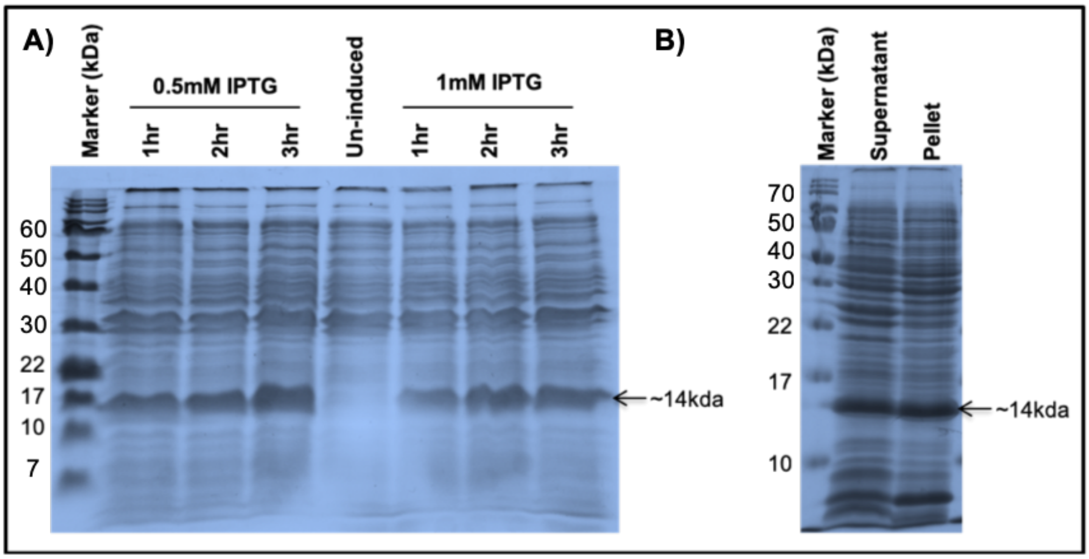


Figure 5.18:- His tagged-Achi DBD induction and fractionation. A) Achi DBD induction was carried out using 0.5mM and 1mM IPTG for 3hrs and in both conditions (0.5mM and 1mM IPTG) showed the successful induction of protein of interest and confirm by the expected size of band ~14kDa B) Achi DBD protein was significantly enriched in soluble fraction of fractionation.

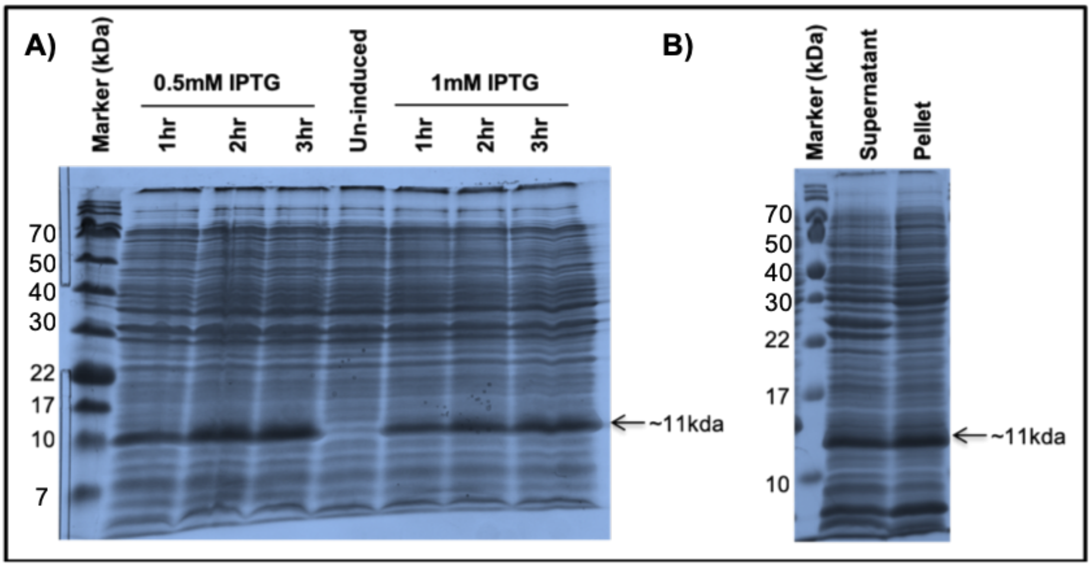


Figure 5.19:- His-tagged Vis DBD protein induction and fractionation. A) Vis DBD protein was induced by 0.5mM and 1mM of IPTG and successful induction of protein was confirmed by expected size of band~ 11kDa. B) Vis DBD was significantly enriched in soluble fraction (supernatant) as well as non-soluble fraction.

5.6. Kr (Kruppel)

The *kr* gene encodes a 502 amino acid multiple Zn finger protein with molecular weight of 54.72kDa (Li et al. 2008) **(Figure 5.20)**. The translated Kruppel protein contains 5 Zn fingers clustered in the central region of the protein (amino acids 222-354) **(Figure 520)**. Binding site for Kruppel protein in the genome is known (Li et al. 2008) and this protein will be used as a positive control for our SELEX experiments.

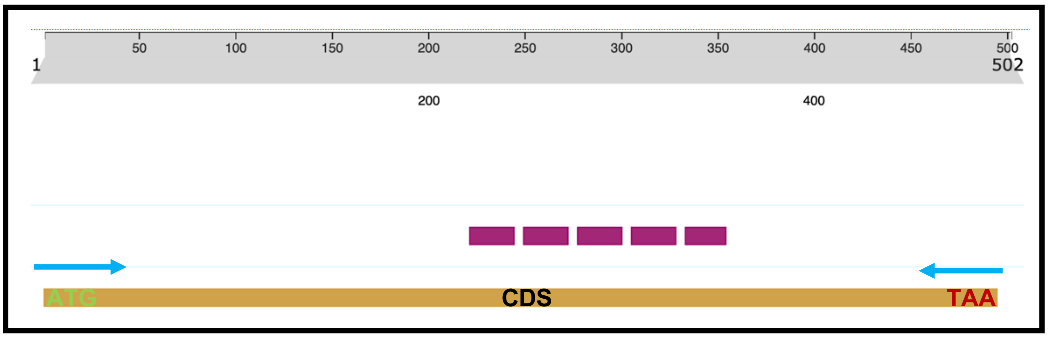


Figure 5.20:- Protein sequence representation for Kruppel and primer design. Kruppel is made up of 502 amino acids and has 5 Zn-fingers (highlighted in purple colour). Primers for amplification of CDS were designed and are shown by blue arrows.

5.6.1. Kruppel FL cloning and protein expression in host *E. coli*

In-order to express Kruppel protein for biochemical characterisation and protein-DNA interaction analysis by SELEX, substantial amount of protein must be produced. Set of primers were designed for Kruppel FL nucleotide sequence **(Figure 5.20)**, the cDNA was amplified using these primers, and PCR product was purified. Purified PCR product was subjected to restriction digestion using appropriate restriction enzymes and cloned in restriction digested expression vector pET28a(+). Cloned insert was confirmed by selecting a number of colonies and performed colony PCR **(Figure 5.21)**. Colony PCR confirmed the positive clones and expected band size was observed **(Figure 5.21)** and cloned sequence of positive clones were sequenced. Sequence confirmed expression vector for Kruppel FL was used for protein expression host *E. coli.* Expression vector was transformed into BL-21(DE3) and allowed to grow on Kanamycin containing LB agar plates for overnight (see Material and Methods). A single colony was picked and incubated in 100ml of LB broth with Kanamycin and protein was induced with 0.5mM or 1mM IPTG. Cells were allowed to grow for 3hrs at 25^{0}C with rotation at 180rpm (see Material and Methods). After incubation, cells were harvested by centrifugation and total lysate was used for SDS-PAGE analysis. An equal concentration of samples were taken and separated by SDS-PAGE and expected band size for induced Kruppel protein was observed (56kDa) **(Figure 5.22A)**. To test the solubility of Kruppel FL protein, the expression vector for Kruppel FL was induced by 1mM IPTG and incubated for 3hrs at 25^{0}C with rotation at 180rpm. After incubation, cells were harvested and lysed with standard lysis buffer (see Material and Methods), and then fractionated by centrifugation. Equal concentration of protein fractions (supernatant and pellet) were taken from the sample and separated by SDS-PAGE. Gel electrophoretic analysis showed that Kruppel FL protein was significantly enriched in supernatant fraction compared to pellet fraction and expected band size for His-tagged version of protein was observed (56kDa) **(Figure 5.22B)**.

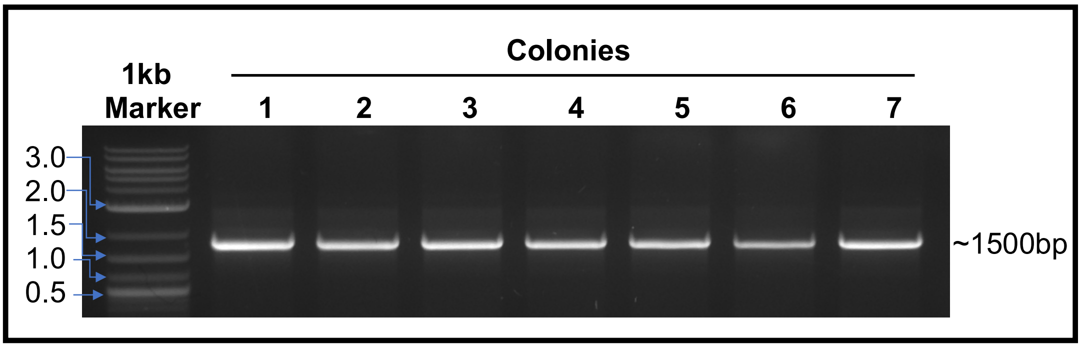


Figure 5.21:- cDNA PCR amplification of Kruppel and colony PCR. To confirm the insert, colony PCR was carried out and expected PCR product was observed for corresponding colonies.

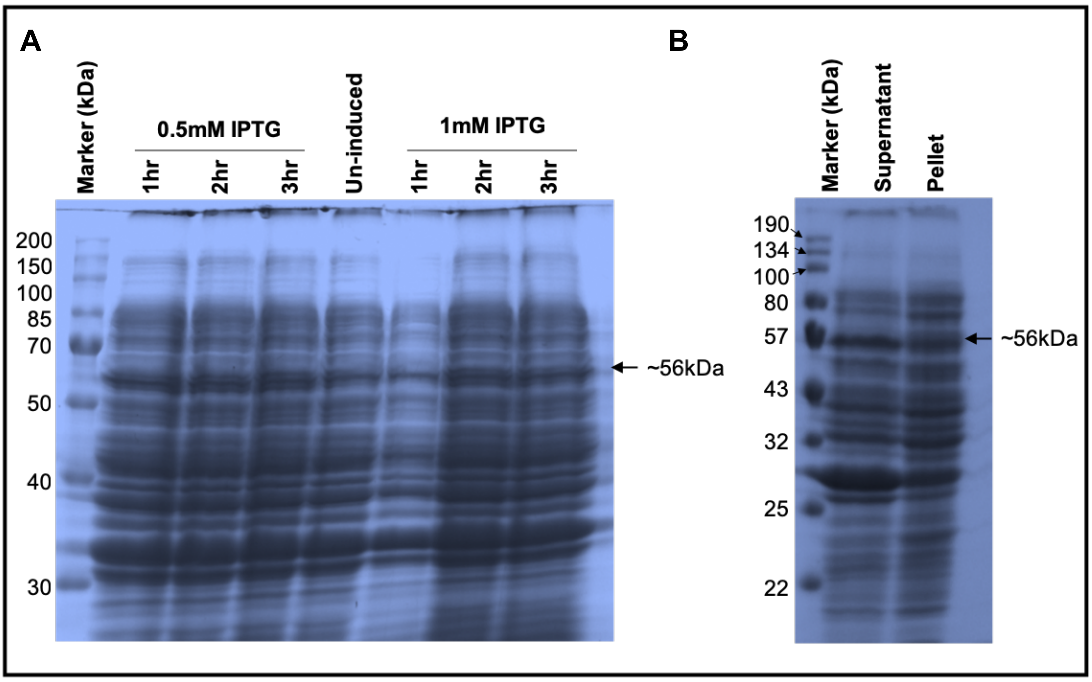


Figure 5.22:- His-tagged Kruppel FL protein induction using IPTG and soluble fractionation. A) Significant amount of Kruppel FL expression was observed in both conditions 0.5mM and 1mM IPTG and expected size of band was observed ~56kDa. B) Kruppel FL protein was present in soluble fraction and expected size of band was observed for respective protein ~56kDa.

5.7. Summary

In-order to study the protein-DNA interaction, recombinant protein of interest must be generated. In this chapter, I had cloned, expressed and fractionated the full length as well as DBD version of four subunits of tMAC complex. C-terminal end of Topi was the first protein to be cloned and expressed amongst other tMAC subunits, and for *in vitro* protein-DNA interaction analysis, I will be using well conserved Zn-finger in C-terminal end of Topi and dsDNA library.

Chapter 6. Results

Identification and characterisation of DNA binding site for protein Topi C-terminal using bead-based SELEX-Sequencing

6.1. Abstract

Transcription factors are responsible for gene expression regulation and discovering their binding sites in the genome of a particular organism is an important question for molecular biologists. To date, various technologies are available to detect the binding sites for the given transcription factors and Systematic evolution of ligands by exponential enrichment (SELEX) (Oliphant, Brandl, and Struhl 1989, Tuerk and Gold 1990) is one of the technologies which allows to identify the in vitro binding site preferred by the transcription factor. A version of SELEX, Bead-based SELEX is a robust method for rapid selection of aptamers (1-6 rounds of selection) and applicable to most targets (small molecules, peptides, protein, and cells) (reviewed in Ozer, Pagano, and Lis 2014).

In this chapter I will present the study of Topi C-terminal (Topi C-term) and its preferred binding sites. The relative molecular weight of His-tagged Topi-C-term is about 37kDa and it has a total of 5 predicted Zn-finger domains. His-tagged Topi C-term protein was allowed to bind to agarose Ni-NTA beads and this complex was incubated with random dsDNA library. Then the protein bound DNA was eluted, amplified and was subjected to the next round of selection. A total of three rounds of selections were performed and DNA from these rounds was subjected to Mi-sequencing along with random dsDNA library. Using motif discovery tools, a 15 bp long *de novo* motifs were discovered from a total of three rounds of SELEX and 7 positive and 3 negative oligos designed based on the discovered motif were tested against the protein Topi C-term. Electrophoretic mobility shift assay suggested that all the 7 positive oligos have affinity towards Topi C-term whereas negative oligos showed no affinity for Topi C-term.

6.2. Characterisation of newly synthesised dsDNA library

In-order to study *in vitro* protein-DNA interaction by SELEX, a dsDNA library must be synthesised and checked for its integrity. The precise design of random fragment library is very important in-order to avoid nucleotide biases. A total of 58 bp hand mixed ssDNA library along with 20 bp random DNA flanking 18 & 20 bp primers (Integrated DNA Technologies) (Murphy et al. 2003) was successfully converted into dsDNA library by Klenow fragment-I. The synthesised dsDNA library was purified by Microcon YM-30 (Millipore) centrifugal filters (Merck) and was checked on 2% agarose gel **(Figure 6.1A)**. Further to check the integrity and concentration of the dsDNA library, it was subjected to Tape-station and single band of expected size-58bp for dsDNA library was observed on virtual gel **(Figure 6.1B)** and graphical representation also showed the single peak of 58bp **(Figure 6.1C)**. The concentration of dsDNA library was 333ng/μl.

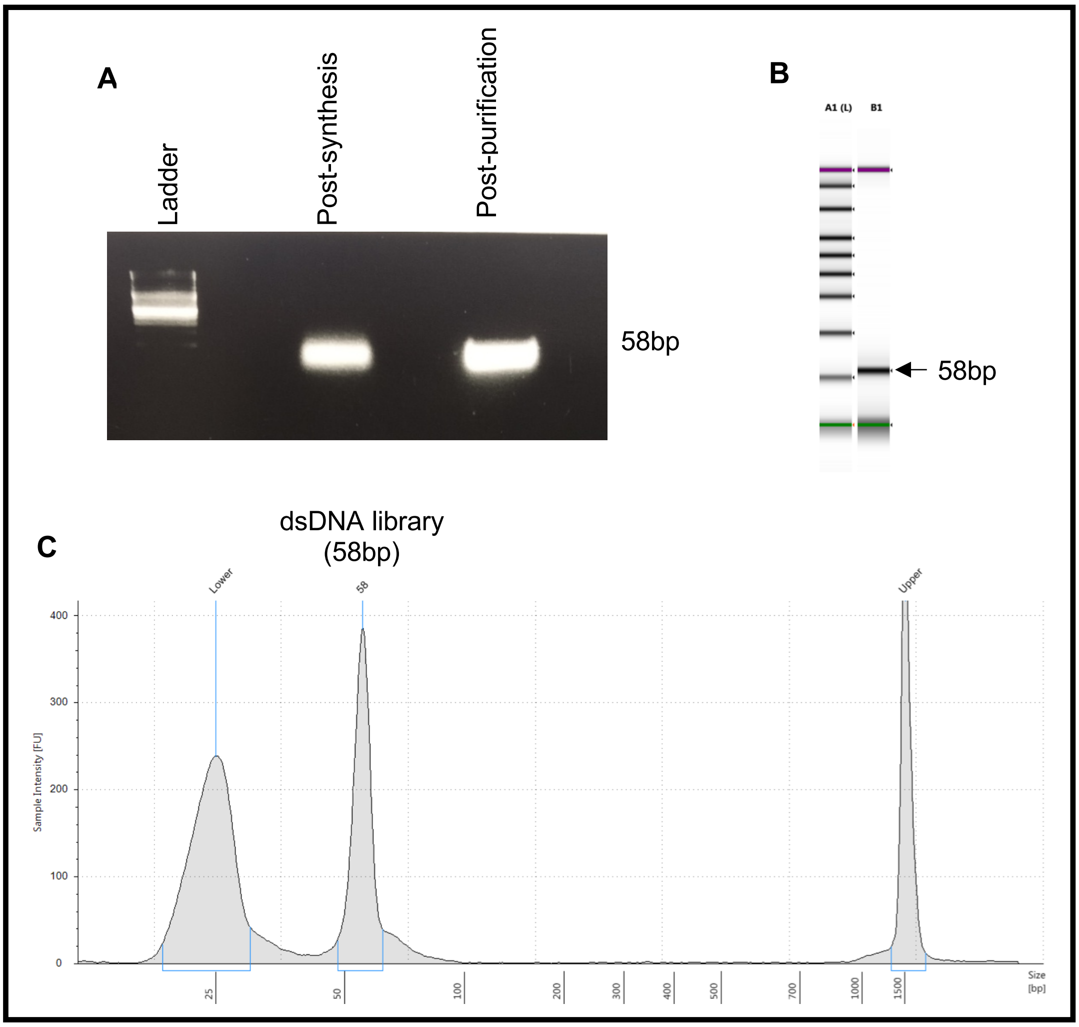


Figure 6.1:- Tape-station analysis of newly synthesised dsDNA library. A) Post-synthesis and post-purification of dsDNA library was checked on agarose gel electrophoresis and single band of 58bp was observed in each well of the gel. B) Virtual gel electrophoresis generated by Tape-station machine showing the size of dsDNA library. C)Tap-station analysis shows that synthesised dsDNA library is the correct size and it is 58bp long. Middle peak represents the length of the library and left and right peaks are lower and upper range of the marker (Agilent High Sensitivity D1000 Ladder) respectively.

6.3. Recombinant protein Topi C-term affinity purification by Ni-NTA agarose beads

For protein expression of Topi C-term, the expression vector was incubated for 3hr in LB with *Kanamycin* and protein was induced by addition of 1mM IPTG **(Chapter 5, Figure 5.3B)**. Cells were harvested and fractionated and enriched protein was observed in supernatant fraction **(Chapter 5, Figure 5.4C)**.

For SELEX, protein-beads complex was prepared and the purity of recombinant His-tagged Topi C-term was confirmed. 1ml of supernatant from fractionation step was incubated with 200µl of Ni-NTA agarose beads and protein allowed to bind to beads for 1hr at 4^{0}C with rotation. Later, beads were washed three times with wash buffer and stored at 4^{0}C until further analysis. 20µl of beads were taken for small scale purification and it was eluted in 20µl of elution buffer. Purity of protein was confirmed by SDS-PAGE **(Figure 6.2)** and expected size of His-tagged version of Topi C-term was observed (~37kDa). The concertation of purified Topi C-term was checked by Bradford reagent and it was 0.3µg/µl. A total of 6 µg of protein yielded from 20µl beads and it was enough for SELEX (for SELEX in 1ml reaction volume, 100pM i.e., 3.7ug of protein was needed).

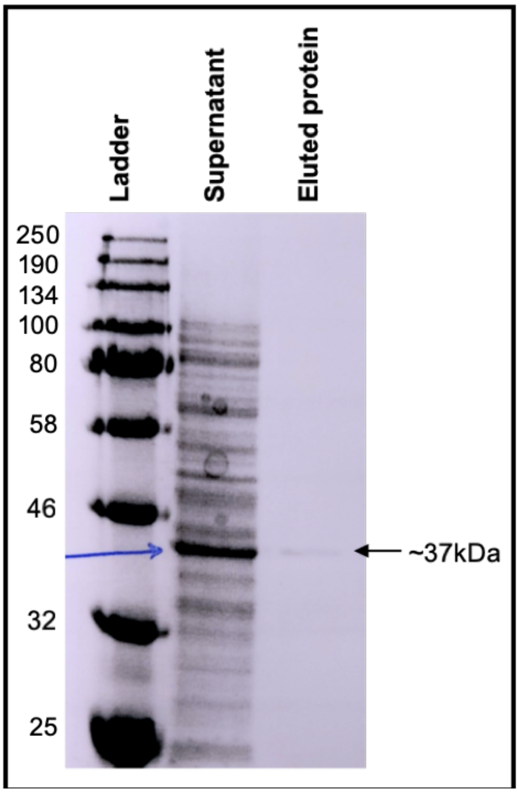


Figure 6.2:- Ni-NTA agarose beads purification of Topi C-term. His-tagged version of Topi C-term was affinity purified from 1ml of supernatant and showed the expected size of band ~37kDa was observed (shown by an arrow).

6.4. SELEX-sequencing discovered *de novo* DNA binding domain for protein Top C-term

A total of three rounds of SELEX were carried out and for the first round, 100pmol of beads bound to Topi C-term was incubated with 1nM of dsDNA library for 30min at room temperature. After washes, protein-DNA complex was eluted with elution buffer and eluted complex was subjected to Emulsion PCR to amplify the eluted DNA (see Material and Methods). After completion of PCR, the concentration of amplified DNA was checked using Qubit by Invitrogen and this newly amplified DNA was used for next round of SELEX. For subsequent rounds of SELEX, 50pmol of beads bound Topi C-term was incubated with 0.5nM of amplified DNA from earlier round. For Nanopore sequencing, 120bp long libraries were prepared by adding adapter sequences to amplified DNA (according to manufactures protocol NEXTflex^{TM} Rapid DNA-Sequencing Kit) from dsDNA library and all the three rounds. The generated libraries were subjected to Next-generation sequencer (Nanopore sequencer) and sequenced from both ends of the libraries. Total number of reads 16000, 30121, 30920, 28447 were obtained from dsDNA library, round1, round2, and round3 respectively. During trimming of the primer sequences, it was found that reads from round-2 and round-3 were contaminated and the contaminated reads were eliminated and only correct reads were taken for analysis. The BLAST search results found that the contaminated reads were from bacteria and human, and it could have happened during library preparation. After trimming, for round-2 and round-3 total 520, 157 reads were retrieved. Trimmed reads from three rounds were uploaded to MEME suite v5.0.1 for motif discovery and dsDNA library was used as a background model (automatically generated by MEME tool).

A motif is an approximate sequence pattern that occurs repeatedly in a group of related sequences. MEME represents motifs as position-dependent letter-probability matrices that describe the probability of each possible letter at each position in the pattern. For motif discovery, Discriminative search mode was used and other parameters were default. The MEME (Bailey et al. 2015) discover motifs that are enriched in first (sequence reads from SELEX as primary) set relative to the second (dsDNA library as control) set. In Discriminative mode, tool first calculate a position-specific prior from two sets for sequences (primary and control). MEME then searches the first set of sequences for motifs using the position-specific prior to inform the search and finds the best motif width. The MEME (Bailey et al. 2015) *de novo* motif discovery is an objective function based on the statistical significance of the log likelihood ratio (LLR) of the events of the motif. The *E*-value of the discovered motif is an estimate of the number of the motifs (with the same width and number of events) that would have equal or higher log likelihood ratio if the input sequences had been generated by random according to the background model (generated using control set of sequences where each position is independent and letters are chosen according to the background letter frequencies).

The MEME usually finds most statistically significant (low E-value) motifs first and It is unusual to consider a motif with an E-value larger than 0.05 significant (Bailey et al. 2015). MEME suite discovered one statistically significant motif from first round of SELEX with *E*-value-7.3e-003 **(Figure 6.3A)**. Four different statistically significant motifs were discovered from second round of SELEX with *E*-value-8.1e-290, 2.4e-026, 1.1e-016, and 9.2e-003 respectively **(Figure 6.3B)** and one motif was discovered from third round of SELEX with *E*-value-1.6e-012 **(Figure 6.3C)**. The overall height of each stack (as in each base at particular position) indicates the sequence conservation at that position (measured in bits), whereas the height of nucleotide symbols within the stack reflects the relative frequency of the nucleotides at that position.

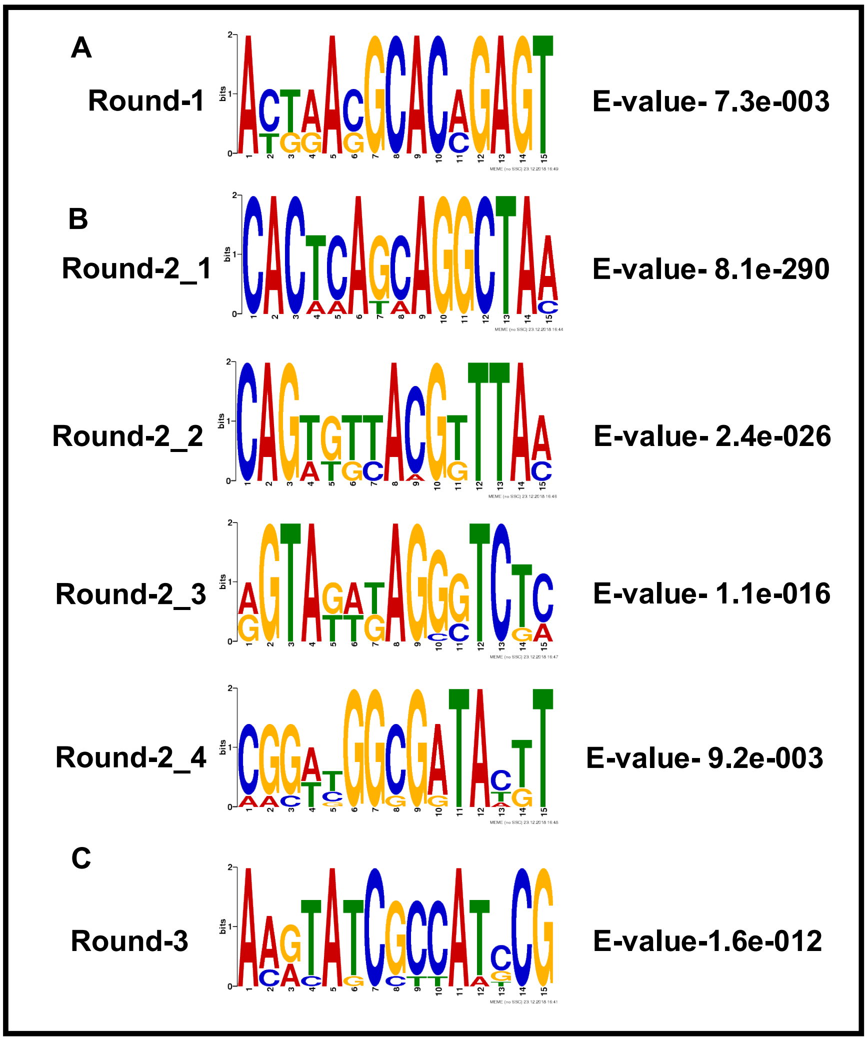


Figure 6.3:- De novo motif discovery for Topi C-term by SELEX-sequencing. Total three rounds of SELEX were performed and 6 different motifs were discovered by MEME suite online tool. A) Motif discovered from first round of SELEX with *E*-value 7.3e-003.B) Total four motifs were found from second round SELEX with *E*-value 8.1e-290, 2.4e-026, 1.1e-016, and 9.2e-003 respectively. C) From round third, one motif was discovered with *E*-value 1.6e-012. Motifs shown here are statistically significant with E-value lower than 0.05.

6.4. Motif comparison found similarities between round 2 and 3 motifs

Motif comparison tells us the similarities between two independent motifs and this could be useful to validate the consistency of the motifs discovered from various rounds of oligo selection. Here, I wanted to test the significant similarity between two newly discovered motif by MEME from round 2 and 3 using TOMTOM by MEME suite using default parameters (Grant, Bailey, and Noble 2011).

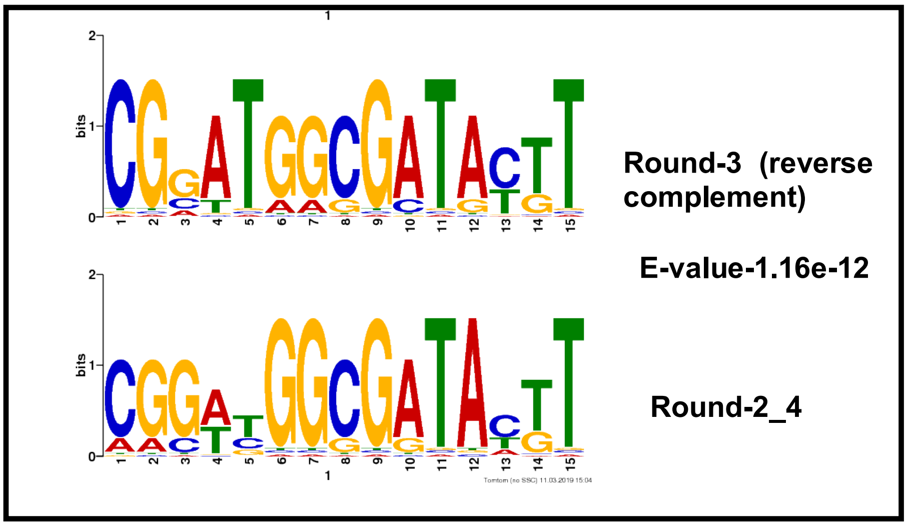


Figure 6.4:- Motif Comparison by TOMTOM between SELEX round 2_4 and round 3. Tomtom algorithm assigns an *E*-value of 1.16e-12 to this particular match.

Tomtom is an algorithm for motif comparison that positions the target motif in a given query motif or in a given database according to the estimated statistical significance of the match between query and the target motifs. Position weigh matrix (PWM) for the motif 2_4 from the round 2 and round 3 were fed in to Tomtom online tool and algorithm assigns as *E*-value of 1.16e-12 for this match between two motifs **(Figure 6.4)**. Alignment between these two motifs (motif 2_4 and motif_3) increases our confidence in motif_3 being real.

6.5. Motif_3 from round-3 occurs in genes expressed in primary spermatocytes dependent and independent of *aly*

FIMO is an online tool that enables to search for given motif in the genome of a specified organism or sequence. Here, I wanted to search for the motif_3 in promoter sequences of genes expressed in primary spermatocytes dependent and independent of *aly*. Over 1000 genes are 16x or more under-expressed in testes mutant for the TMAC subunit, *aly*, compared to control (Doggett et al. 2011). I chose motif_3 because it shows good match with motif 2_4 from the round 2, hence suggest that motif_3 could be real **(Figure 6.4)**. Testis-specific promoters are typically short: 200bp, or even less, of 5' genomic flanking sequence is usually sufficient to confer testis-specific (White-Cooper 2010). The region flanking the annotated (TSS ± 250bp) for genes dependent and independent of *aly* were downloaded from Flybase database and identifiers for each sequence were modified according to FIMO requirements. Overall, FIMO identified 85 sites matching motif_3 identified by SELEX in promoter regions of genes in *aly* mutant background and 21 in wild types using *p*-value-<0.0001. With *p*-value-<0.001, a total of 838 potential binding sites were found in promoter regions of *aly* dependent genes and 180 in *aly* independent genes. Further with the *p*-value-<0.01, a total of 8860 candidate binding sites were found in promoter regions of *aly* dependent genes and total 2179 potential binding sites were found in promoter regions of *aly* independent genes **(Figure 6.5)**. The FIMO analysis suggested that there are higher number of potential binding sites for Topi C-term in the promoter regions of *aly* dependent genes compared to *aly* independent genes. This increases our confidence in discovered motif being real.

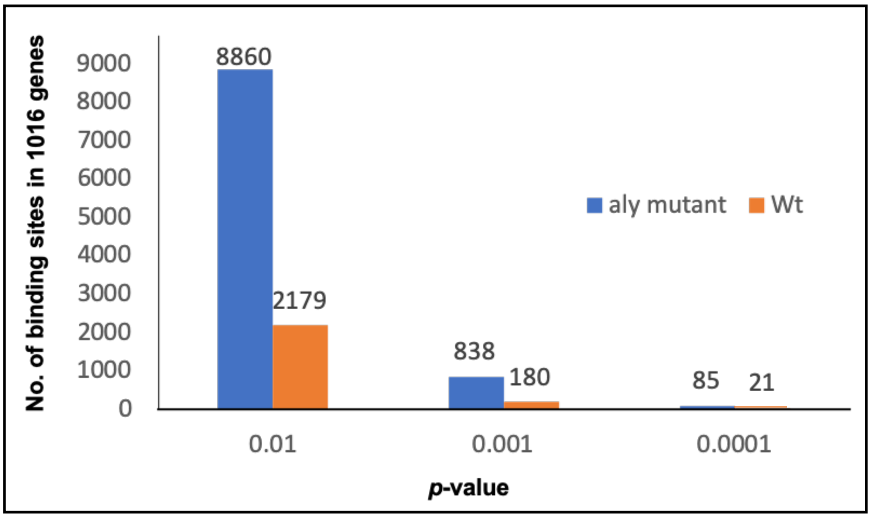


Figure 6.5.:- Occurrence of motif_3 in promoter regions of *aly* dependent and *aly* independent genes. Graph represents the motif occurrence using default parameters (*p*-value-0.0001) and lower *p*-value for search (*p*-value 0.001, 0.01).

6.6. Motif_3 aligns with the predicted motif for C-terminal end of Topi

Topi FL protein has predicted 11 *Zn* fingers spread across the length of the protein sequence. The C-terminal end of the protein has total 5 Zn-fingers and N-terminal end has 6 Zn-fingers. Here, I wanted to test all the discovered motifs for Topi C-term with the predicted motif for Topi. For that, the PWM for Topi FL protein was generated by feeding the protein sequence to online tool (http://zf.princeton.edu/).

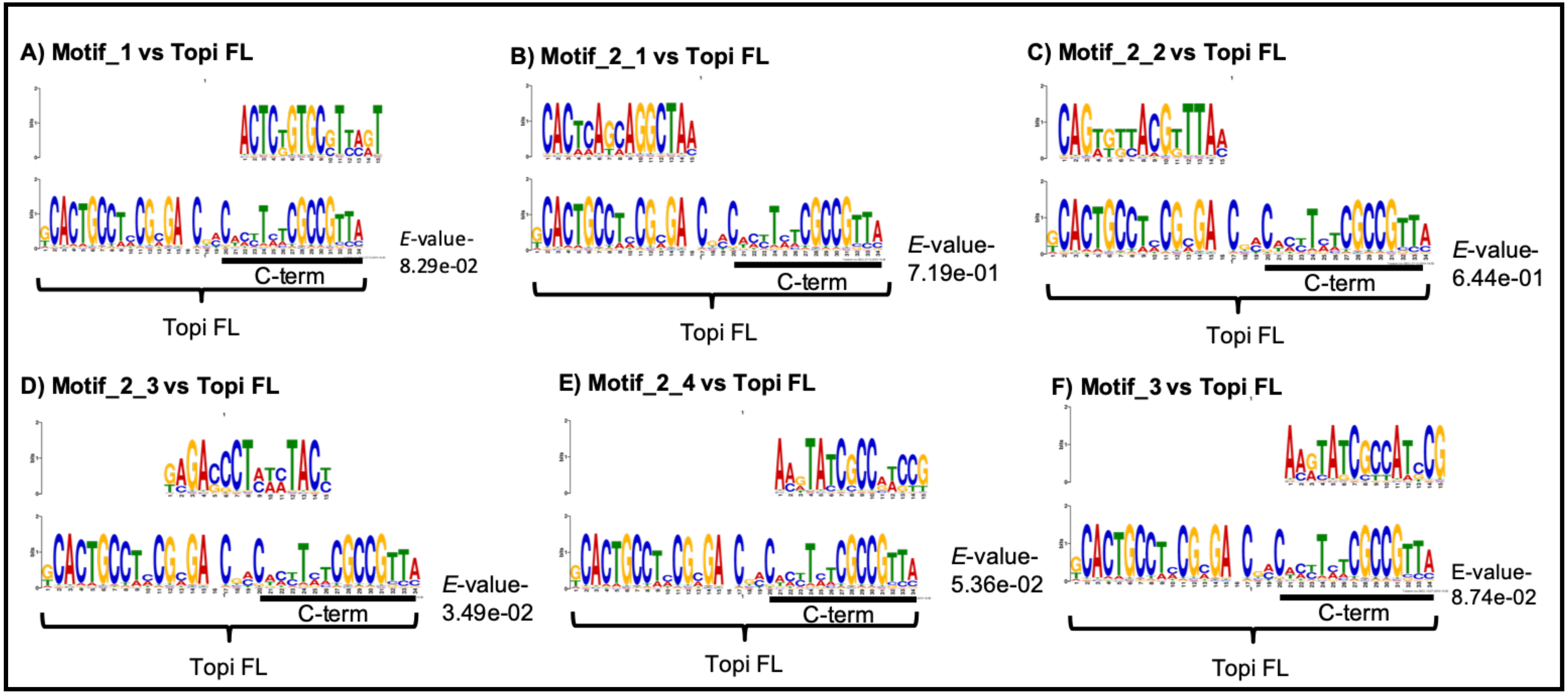


Figure 6.6:- Motif comparison between PWM of motif discovered from all the three rounds of SELEX and predicted motif for Topi FL by online tool. Alignments generated by TOMTOM for A) Motif_1 vs Topi FL, B) Motif_2_1 vs Topi Fl, C) Motif_2_2 vs Topi FL, D) Motif_2_3 vs Topi FL, E) Motif_2_4 vs Topi FL, and F) Motif_3 vs Topi FL.

The PWM for motif_1, motif_2_1, 2_2, 2_3, 2_4, and motif_3 and PWM generated by the online tool for Topi FL protein were fed to Tomtom and algorithm assign an *E*-value of 8.29e-02, 7.19e-01, 6.44e-01, 3.49e-02, 5.36e-02, and 6.92e-02 respectively for generated alignments **(Figure 6.6 A-F)**.

6.7. Characterisation of motif_3 by Electrophoretic mobility shift assay(EMSA)

SELEX sequencing discovered motifs for C-terminal end of protein Topi. The motif_2_4 and motif_3 share similarities and based on these two discovered motifs, we designed 15 bp long, 7 different non-labelled probes and 3 negative probes, and 4 additional bases were added on both sides of the probes in order to allow protein to bind to the probes. Positive probes were designed based on motif 2_4 and motif_3 with different combinations and a nucleotide which is highly conserved (with higher bit score) at given position in motif was kept the same in the probe sequence. Negative probes have the similar AT:GC ratio as the positive probes considering these probes would not have any affinity towards C-terminal end of protein Topi. To generate dsDNA probes to use in *in vitro* binding assays two single stranded complementary oligonucleotides were synthesised. Equimolar ssDNA probes were mixed with its complementary ssDNA probes and renatured by incubating at 95^{0}C for 5 min cooled down slowly. Each dsDNA positive and negative probe was allowed to bind to purified Topi C-term **(Figure 6.7)** and whole complex was resolved on 2% agarose gel in 0.5xTBE buffer.

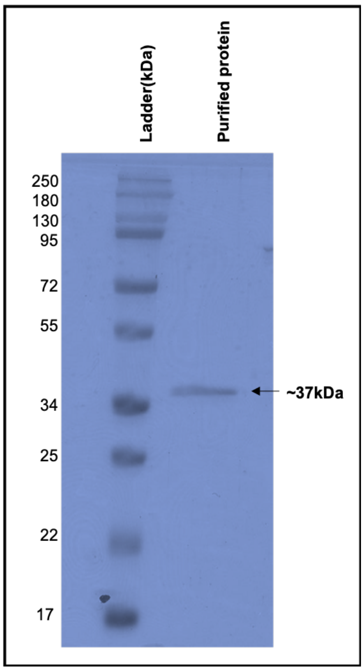


Figure 6.7:- Ni-NTA beads purification of Topi C-term. Topi C-term protein was purified by Ni-NTA agarose beads with standard buffer provided by Qiagen.

A shifted band produced by non-labelled positive dsDNA probes bound by C-term Topi could be directly visualized on UV transilluminator after EMSA-agarose gel electrophoresis **(Figure 6.8)**. Positive dsDNA probes appeared to be bound by C-term of Topi and protein-DNA complex and unbound dsDNA probes can be visualize in gel whereas none of the negative dsDNA probes were bound by C-term of Topi **(Figure 6.8)**. Also, in the absence of protein in reaction, free probes appeared to be mobilizing in gel at the same rate as unbound probes were observed in protein-DNA reaction **(Figure 6.8).**

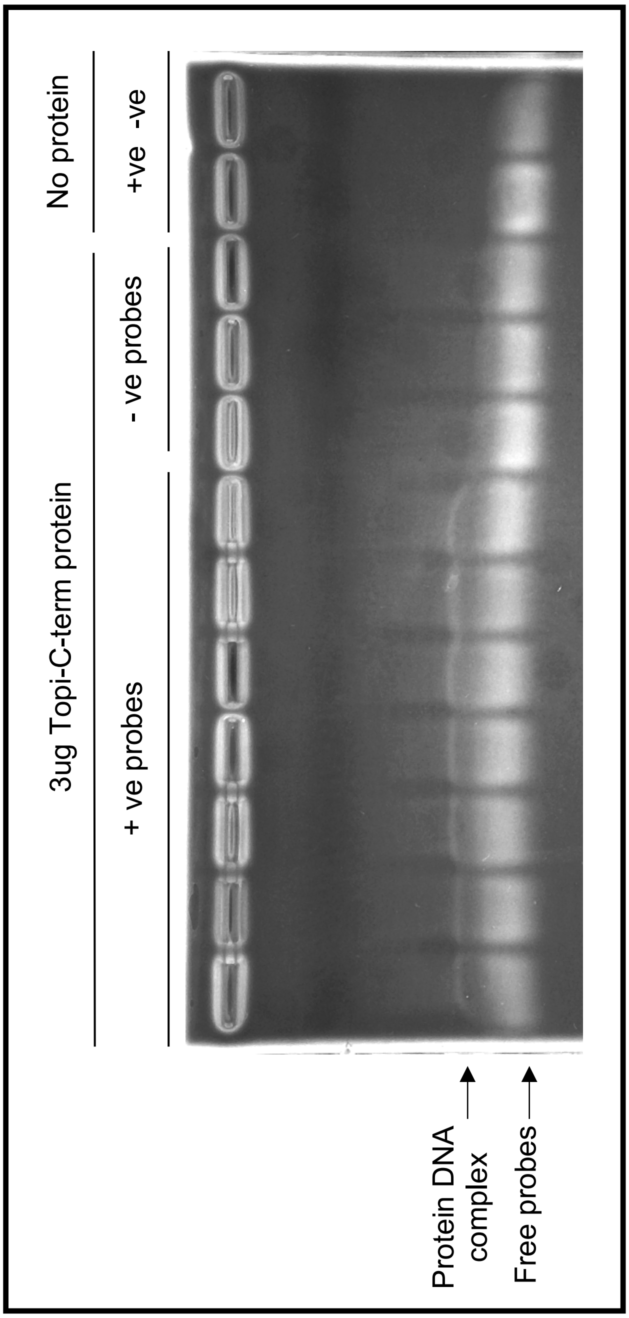


Figure 6.8:- Electrophoretic Mobility Shift Assay using protein Topi C-term with oligos designed from motif discovered by *in silico* analysis. 3ug of protein was incubated with equimolar concentration of individual positive probes as well as negative probes and mobility shift was only observed in positive probes whereas negative probes showed no shift. An individual mixture of positive and negative oligos observed as free probes in the absence protein Topi-C-term.

6.8. Summary

SELEX-sequencing is a robust *in vitro* method for selection of oligos against specific ligand molecule of interest. Our results suggest that three rounds of selection of SELEX discovered the putative binding site for C-terminal end of Topi. Discovered motif_3 aligned with the C-terminal end of the predicted motif of Topi FL protein. EMSA results suggest that designed positive probes show affinity towards purified protein C-terminal of Topi whereas negative probes show no affinity whatsoever. Discovered motif was also compared with other know Zn finger proteins and good alignment was observed for tested proteins. Topi paralogue Kruppel also showed the alignment with the discovered motif for Topi C-term (data not shown). These all various analyses raises our confidence in the motif_3 being real.

Chapter 7. Discussion

7.1. Functional analysis of promoter region of *de novo* genes

Drosophila spermatogenesis is a unique process characterized by cellular differentiation encompassing three main phases namely (i) mitosis, producing more spermatogonial cells, (ii) meiosis, production of genetically diverse haploid cells and (iii) spermiogenesis, involving maturation of spermatids to sperms that comprises nuclear restructuring and chromatin packing (amongst many other morphological changes). Spermatogenesis has become a paradigmatic system to study gene expression regulation because of the specialised transcriptional event occur during the transition of spermatogonia to spermatocyte (White-Cooper 2010, Gan et al. 2010, Ayyar et al. 2003, Jiang et al. 2018, Jiang and White-Cooper 2003).

An increasing number of reports have shown the expression of *de novo* genes in various organisms and making it interesting to study the gene expression regulation of these *de novo* genes (Begun et al. 2007, Cai et al. 2008, Carvunis et al. 2012, Knowles and McLysaght 2009, Neme and Tautz 2013, Reinhardt et al. 2013, Ruiz-Orera et al. 2015, Schlötterer 2015, Witt et al. 2019, Xiao et al. 2009, Zhao et al. 2014, Gubala et al. 2017). *de novo* genes evolve from previously non-genic regions. It is an important evolutionary question to address what mutations are implicated in this process. Analysis of the sequences can tell us what changes have occurred, but only a detailed molecular analysis can determine which of the changes are causative to the gain of gene expression. This is the first study to discover the molecular mechanism of the promoter region of *de novo* genes in *Drosophila*. Our findings provide an unprecedented perspective on molecular evolution of two testis-specifically expressed *de novo* genes. In *Drosophila*, *de novo* genes are evolved from non-coding DNA sequences and genes I studied are still segregating. It is not clear if they will become fixed or if they will be lost from the population (Zhao et al. 2014). *de novo* genes are known to be expressed at low levels especially in testis of various organisms for example fruit flies, humans, chimpanzee, rodents (Zhao et al. 2014, Knowles and McLysaght 2009, Ruiz-Orera et al. 2015).

7.1.1. Origin and spread of *de novo* genes

In general, origin of *de novo* genes and how they are spreading in the population is not very well known. Although, experimental identification of *de novo* genes depends on comparison between the species of the interest and a closely-related sibling species and this comparison yields orthologous DNA sequences. For example, *de novo* gene *Polymorphic derived intron-containing* (*Poldi*) was found in *Mus musculus* and other very closely related species that has occurred in the last 2.5-3.5 million years (Heinen et al. 2009). *Poldi* is expressed in testis, at the round spermatid stage, and knocking out of the genomic region leads to a reduction in male fertility, indicating that gene has attained a biological function. Comparison of the *Poldi* genomic region between species in which the gene is expressed and those in which there is no expression shown a small number of derived sequence changes in the promoter and just the downstream of the TSS that correlates with the expression/non-expression status. For example, expressing strains have an 11 bp deletion relative to non-expressing status at 390 bp.

7.1.2. cis-acting regulation of de novo genes

In *Drosophila*, approximately one-third of the *de novo* genes are expressed in at least one strain and can be identified by the presence of SNP within 500bp upstream of the TSS that correlates with a difference in gene expression (Zhao et al. 2014). This supports the idea that single nucleotide change can be responsible for initiation of transcription within an intergenic sequence. Furthermore, in *Drosophila,* many segregating *de novo* genes are dependent on the cis-acting regulatory elements for their expression (Zhao et al. 2014). Moreover, these studies also found four common 8- and 10-bp consensus motifs in -100bp to +50bp region and 23% annotated male biased genes also share these motifs and suggest segregating *de novo* genes is mainly influenced by cis-acting elements and it was tested directly by looking at heterozygotes with one expressing allele and one non-expressing allele. The authors determined that in almost all the cases the expressed allele was expressed, and the non-expressed was silent in these heterozygotes, therefore the genetic control of the allele expression had to be cis-acting (Zhao et al. 2014). If expression of the allele was driven by the trans-acting factors, then both parental allele would express in heterozygotes and a gene whose expression was affected by cis and trans variants would express both parental alleles in the F1.

7.1.3. 7 bp Indel-I is necessary and sufficient to drive gene_090 expression in testis

Gene_090 is a segregating *de novo* gene and is only expressed in *Drosophila* testis (Zhao et al. 2014, Witt et al. 2019). The expressed allele and non-expressed allele differ by 6 SNPs and 2 indels. In chapter 3, I have demonstrated that the various reporter constructs made with the promoter region of gene_090 is capable of regulating the transcription and protein is expressed in post-meiosis.

Our studies have established that difference between expressed and non-expressed allele of gene_090 can be attributed to sequence differences with this short yet important

DNA sequence. Our results show that presence of the SNPs in promoter region of the gene_090 are not important to gain expression in testis. It would be interesting to study the promoter regions of the other segregating *de novo* genes by designing and testing the synthetic DNA constructs in reporter construct. The cis-regulation of testis-specific transcription is little known phenomena. A few studies on several genes indicate that short genomic region encompassing the TSS are sufficient to gain testis-specific transcription and combined activities of regulatory elements within these genomic regions provide testis specificity and regulate transcription level (White-Cooper 2010). For example, genomic region comprising from -53 to +172 bp of the *beta 2 tubulin* gene is sufficient to drive the expression in testis (Michiels et al. 1989, Santel et al. 2000). Also, the beta2UE1 element at -45 bp is crucial for testis-specific transcription and for high level of transcription, two elements beta2UE2 at -29 bp and beta2DE1 at +60 bp are necessary.

Bioinformatic analyses on 190 genes expressed in testis identified a 10 bp A/T-rich motif that is identical to the TCE (Katzenberger et al. 2012). This study showed that TCE can be important for getting high levels of transcription, but that it alone is not responsible for getting the testis expression. Our findings are consistent with earlier studies and suggest that TCE site in the expressed allele does not act as TCE and poses more questions about transcriptional regulation of *de novo* genes. In the light of unusually short transcriptional regulatory elements, gene expression might be regulated by cell type-specific TFIID subunits (Katzenberger et al. 2012).

Through analysing the promoter region of gene_090 in *LacZ* reporter construct, we have showed that the expression of gene_090 is regulated by the regions upstream of the TSS. The activating deletion found in series of experiments suggest that new promoters in *de novo* genes could emerge by creating new binding site for regulatory proteins. As a future perspective, it would be interesting to find the regulatory proteins involved in regulation of *de novo* genes and it will help address the mechanism of *de novo* gene expression regulation. The potential regulatory proteins could be discovered using DNA pull-down assay followed by mass spectrometry in which biotin labelled DNA probe would be incubated with cell lysate. Then, the protein-DNA complex would be purified using anti-biotin antibody and protein would be detected by western blot or identified by mass spectrometry. A recent report has shown the requirement of the minimal promoter to drive the expression of lac operon in *E. coli* (Yona, Alm, and Gore 2018). In this study, authors designed 40 random sequences of 103 bp long as the same length of wild type (WT) lac intergenic region and it was replaced with these random sequences. To comply with the GC content of the *E. coli* genome (50.8%), random sequences with GC content (lower than 45.6% or higher than 56%) were omitted. Also, authors excluded the random sequences with homo-nucleotide stretches longer than five to avoid issues with sequencing. The authors showed that a single nucleotide change in computationally generated random sequences (the same way experimental were generated) could produce the same expression levels as wild type lac promoter. In conclusion, this study suggest that minimal mutation in the random sequences could turn these random sequences into functional promoters (Yona, Alm, and Gore 2018). This study was performed in *E. coli,* it may not be the same case in complex eukaryotes. Comparative genomic studies in eukaryotes provide evidence for the evolutionary dynamics of TF binding, highlighting the possibility for rapid and flexible TFBS gain and loss between closely related species on time scale of as little as few million years (Villar et al, 2014). On the other hand, there are examples of rapid evolution of TFBSs across and within population requiring shorter timescales i.e., 10,000-100,000 years (Zheng et al 2011, Contente et al, 2002).

Our results also suggest that more than one protein is involved in expression of gene_090. Based on these results we proposed a model representing gene expression regulation of gene_090. In expressed allele, a 7 bp deletion brings flanking sites together, thus allow regulatory proteins to bind to this site and drive the gene expression in testis **(Figure 7.1A&B)**. Further constructs were consistent with proposed model and suggested that flanking sites of the deletion are also required for expression, hence suggest that more than one protein could be required for gene regulation. Additional constructs showed that adjoining nucleotides of the deletion flanking sites are important for expression too **(Figure 7.1C)**, we conclude that the site bound by one or other of the proteins extends to the region in the middle of the indel. Our hypothesis about creating regulatory protein binding site in the expressed allele suggests that the sequence specificity is also important to drive expression.

The reporter gene is translated in spermatids suggesting that gene has been transcribed in earlier stages of the spermatogenesis. This hypothesis of transcription can be tested by reporter gene mRNA *in situ* hybridisation and this study would help us understand expression pattern. Also, in our lab we have showed that gene_090 requires tMAC for its expression (unpublished data) and studying the expression pattern of gene_090 in tMAC mutant background would help us understand the role of tMAC in *de novo* gene expression regulation. The role of tMAC would be understood by SELEX-seq or ChIP-seq using each DBD containing protein and identify the tMAC binding sites in the promoter region of gene_090.

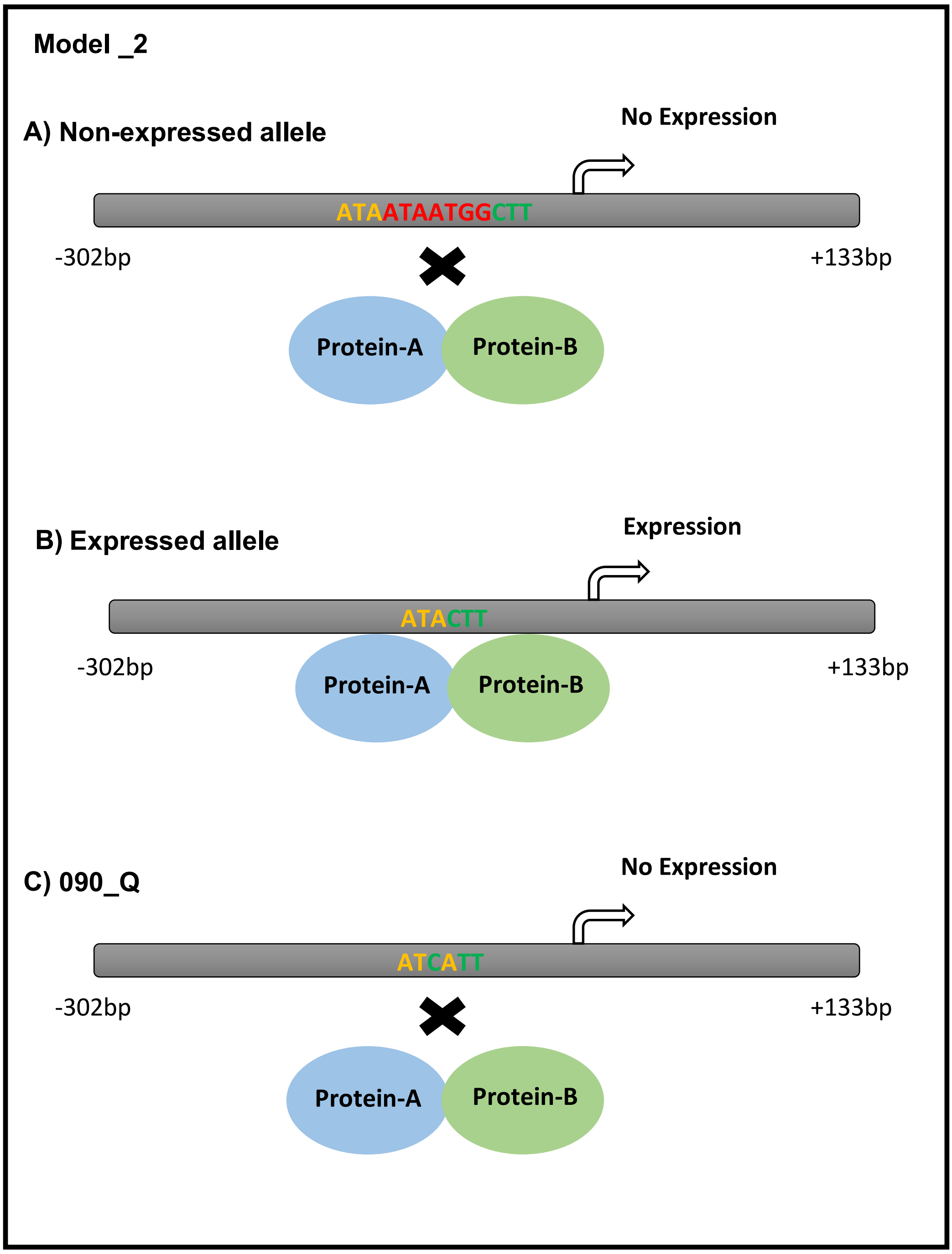


Figure 7.1:- Proposed model representing the mechanism of gene expression of gene_090. A) In non-expressed allele, indel-I (insertion highlighted in red colour alphabets) prevents regulatory protein binding to indel-I site, thus gives no expression whereas B) in expressed allele, deletion of indel-I allow flanking sites to come together and regulatory proteins drives the expression. C) When we have a expressed allele sequence with two bases switched at flanking sites of the indel-I gives no expression in testis. Flanking sites of the deletion are important for expression and it suggest that more than one protein is involved in gene expression regulation.

Gene regulation might be modulated by cell-specific TFIID and a distinct transcription initiation complex characterize mRNAs that will be subjected to translational repression (Blümer et al. 2002). In *Drosophila,* translational repression elements (TREs) (Blümer et al. 2002) are usually located in the 5'UTR and only found in testis-specifically expressed genes. The reporter construct used in our studies has 5' UTR from *adh* gene and other constructs that have *adh* 5' UTR does not cause translational repression. For example, the pCaSpeR-AUG-bGal-*djl*Δ52 construct has a 5' UTR form *adh* gene, but is translated in primary spermatocytes. This suggests that there is no translation repression of reporter gene in *Drosophila* testis.

7.1.4. Gene_074 expression is not associated with SNPs or indels

As explained in chapter 4, expression of gene_074 does not correlate with any SNPs or indels. The expressed allele 517 differs by 1 SNP relative to the non-expressed allele tested and the most likely candidate is the "G" very close to the TSS which is not found in any other alleles. The allele on the *aly* chromosome differs from the non-expressed allele by 8 SNPs and 2 indels. One SNP just upstream of TSS and indels are also in the 517 allele, however this is also in other non-expressed alleles, so this is unlikely to be causative on its own for expression. However, allele of gene_074 from fully sequenced *aly* chromosome and another expressing allele of gene_074 517 differ by 3 SNPs and 1 indels. It is interesting that there is no single SNP or indel that differentiates expressing from non-expressing alleles. This suggests that the expression may have evolved independently in these two alleles.

In our lab we have found that the aly allele of 074 expresses independently of tMAC. but since the 517 allele has evolved separately, and different SNPs are implicated, we don't know yet if that version of the promoter is similarly independent of tMAC for expression. However, it would be interesting to see the RNA expression and localisation of reporter gene in wild type and *aly* mutant background. We can also check whether other components of tMAC complex play any role in the regulation of gene expression. In addition to above, RAL-357 and allele from *aly* chromosome only differ by 1 SNP and share common unknown ancestor. Testing this one SNP could tell us about the expression/no-expression status of these two alleles. RAL 399 is another closely related allele to the allele from *aly* chromosome and these two alleles differ by 3 SNPs and 2 indels. With combinations of SNPs and indels we can design few more constructs to check the difference between these two alleles and it also will tell us about evolution of gene_074

Evidently, reporter gene is translated in spermatids and suggests reporter gene mRNA must be generated in earlier stages of the spermatogenesis like many other testis specific genes (White-Cooper 2010). RNA *in situ* hybridisation against the reporter gene would tell us the mRNA localisation in spermatogenic cells. Also, studying the transcription of reporter gene in tMAC mutants would help us understand the role of tMAC in gene expression regulation. We can also test these construct in other tMAC components, tTAFs, and other meiotic arrest mutants. Also comparison of the aly allele and the 517 allele would be interesting as they may have different mechanism of regulation or they could be transcribed in somewhat different patterns.

7.2. Recombinant expression of proteins: Topi, Comr, Tomb, and Achi/Vis

In primary spermatocytes, tMAC is a transcriptional complex which regulates the expression of more than 1000 genes (White-Cooper 2010, Jiang and White-Cooper 2003). To date, this complex has a total of eight components and four of them have predicted DNA binding domain (Doggett et al. 2011, Jiang et al. 2007, Perezgasga et al. 2004, Laktionov et al. 2014). How tMAC complex regulates gene expression in testis is not well known. It would be intriguing and biologically important to address the mechanism of gene expression manifested by this very complex.

For expression of recombinant proteins, *E. coli* has always been the preferred microbial cell factory. *E. coli* is a suitable host for expressing the eukaryotic proteins. Large-scale production of protein trials have shown that <15% of non-bacterial proteins can be expressed in *E. coli* in a soluble fraction (Braun and LaBaer 2003). Recombinantly expressed full length version of these four subunits of the tMAC complex would be used for *in vitro* identification of their respective binding sites. I also have managed to clone and express DNA binding domains (DBDs) of all four proteins and these would be useful to improve the stringency of our analysis by allowing us to compare between two motifs derived from full length and DBD versions of protein of interest. In the future with help of protein modelling we will deduce the relevant sites in protein with affinity for DNA and we will use these DBDs by making the mutants for relevant interacting amino acids in given protein and retest the motif discovered with given wild type protein. This would tell us the authenticity of discovered motif for protein of interest as well as it would tell us if mutation in DBD of given protein allow tMAC complex to form at target promoter and regulate gene expression. We can also understand the role of tMAC in gene expression by generating the targeted mutants that would retain the ability to form the whole complex but would lack the ability to interact with target DNA regions. In more ambitious plans, we also can express all the components of the tMAC complex and assemble these components in *in vitro*. This approach has already been used and it would be useful for our study to analyse *in vitro* tMAC-DNA interaction using single-particle cryo-electron microscopy. For example, in recent study, human TFIID, Pol II, and TFIIH along with other transcription factor associated with TFIID were immunoprecipitated and recombinantly expressed and purified from *E. coli.* These components were assembled with target DNA sequence in processed assembly buffer. After processing and incubation, this *in vitro* assembled complex was analysed by Cryo-EM and acquired data was processed and three-dimensional protein-DNA complex was generated (Louder et al. 2016).

For protein Achi/Vis, motif identification was carried out via bacterial one-hybrid system (Noyes et al. 2008). For this study, authors mainly focused on TFs which are expressed in early anterior-posterior patterning in *Drosophila.* This network involves TFs of the five most highly represented DNA-binding domain families; C2H2 zinc fingers, homeodomain, bHLH, bZIP and winged helix, and other less well-represented domains. 80% of the sequence-specific TFs utilize one of the DNA-binding domains represented in this group (Adryan and Teichmann 2006) and some of these TFs, such as Kruppel, have very well-defined DNA-binding specificities (Noyes et al. 2008).

In silico analysis by FIMO failed to find the Achi/Vis motif in tMAC target genes with p-value <0.0001. But there were 9516 motif occurrences with a p-value <0.01 and 762 motif occurrences with a p-value < 0.001 discovered. This analysis suggests that motif-3 was enriched in testis-specific promoters if we lower the p-value. By performing the SELEX using Achi/Vis protein we are not only going to verify the published motif, but it also gives an opportunity to identify significantly lower affinity sites, which are likely to be biologically relevant. SELEX coupled with other TFBS finding methods like EMSA, would allow us to select aptamers against full length as well DBD versions of protein of interest and will help us to look at different binding affinities. Weak binding matches to the consensus found during our analysis might be biologically important because as I explained earlier tMAC complex has multiple predicted DNA binding proteins and bioinformatics has failed to find any consensus sequence. This suggests that tMAC complex might bind to promoters of target genes without sequence preferences, thus sequences with weak matches to the consensus suggests to be biologically very important.

7.3. *de novo* motif discovery for C-terminal end of Topi

Many studies have showed that primary TF-DNA binding sites evolve slowly, and is conserved between species (Bohmann et al. 1987, Wei et al. 2010, Amoutzias et al. 2006). Many differences in the TF repertoire between invertebrates and vertebrates exist because of the expansion of some gene families, such as C2H2 zinc finger factors and nuclear receptor in vertebrates (Charoensawan, Wilson, and Teichmann 2010). At present, DNA binding sites of many TFs can be found in many TF-related databases such as TRANSFAC, TFDB, JASPAR, MAPPER, RedFly etc, (Khan et al. 2018, Fornes et al. 2019, Gallo et al. 2011, Wingender et al. 1996). Binding specificities for TFs across various taxa has been determined with different methods like, ChIP (Chromatin Immunoprecipitation), SELEX, EMSA, One-hybrid system, DNase I footprinting, and protein binding microarrays.

In chapter 5 I have explained the recombinant expression and purification of Topi C-term and in chapter 6 I have explained the use of SELEX to identify the *de novo* motif for Topi C-term protein. To date, many studies have shown the selection of aptamers against the agarose or magnetic Ni-NTA beads bound target molecule. A study showed the selection of aptamers against the TTF1 (a member of NK homeodomain transcription factors) target protein by SELEX (Murphy et al. 2003) in which his-tagged protein of interest was allowed to bind to magnetic beads and random library of aptamers was allowed to bind to beads-protein complex. Our results shown that use of Ni-NTA agarose beads to immobilize the protein of interest during the selection. The use of His-tag for immobilization of protein of interest promotes the proper orientation of proteins on a bead surface, and concurrently provides a purification step, hence reduces the chance of selection of contaminants (Murphy et al. 2003).

In this study, we adopted SELEX coupled with high throughput sequencing. This strategy was mainly featured with the following technical improvements. Firstly, we designed the non-labelled hand-mixed ssDNA oligos. Secondly, this strategy introduced a use of Ni-NTA agarose beads to immobilize the protein. Thirdly, we adopted emulsion PCR (Shao et al. 2011) to amplify the DNA from each round. Fourthly, we used non-labelled positive and negative oligos to characterize the identified motif by electrophoretic mobility shift assay (EMSA). The use of hand-mixed ssDNA oligos allowed us to minimize the biases during the synthesis of oligos. Hand-mixing of bases allows to customize the base ratios during synthesis, which includes non-equimolar ratios. In addition, during the library synthesis base ratios are adjusted to balance the slight coupling variations in each base. His-tagged C-term of Topi was immobilized on Ni-NTA agarose beads and purity was checked by SDS-PAGE. SELEX selection were monitored by PCR amplification and quantification of amplified product. Emulsion PCR allowed to amplify the DNA from each round without accumulating biased PCR products. In emulsion PCR complex, millions of cell-like compartments are separated from each other without exchange of any macromolecules or PCR products. A study demonstrated that emulsion PCR is useful in SELEX experiments for unbiased PCR amplification (Schütze et al. 2011). Currently for SELEX experiments most diverse libraries with high variability in nucleic acid pool are used. The emulsion PCR allows to amplify these libraries reduces the by-product or biased products accumulation in reaction and maintains the complexity of libraries. A total three rounds of SELEX selection coupled with emulsion PCR allowed us to identify the motif for Topi C-term.

Motif derived from the round-2 and -3 showed good alignment during our motif comparison analysis using TOMTOM online tool and it suggests that Topi C-term could have relatively strong DNA sequence affinity. PWM for full length Topi protein was generated using online Zn finger prediction tool (Persikov et al. 2014). A motif from round-3 also showed good alignment with the PWM for full length Topi. I have analysed the occurrence of this motif-3 with FIMO in a set of 1016 tMAC target gene promoters (500 bp flanking TSSs, downloaded from Flybase), and found 85 occurrences with a *p*-value <0.0001 (838 at *p*<0.001). Only 21 occurrences were found in a control gene test (testis-expressed, tMAC-independent) at p<0.0001. This raises our confidence in the motif being real, however a strong match is still only present in ~6% of target genes (some genes have multiple sites). When we performed the motif enrichment analysis, we discovered that motif-3 was not enriched at any locus in the given set of sequences but it was dispersed across the sequences.

I also found that Topi C-term motif shares somewhat similar binding site with quite closely related Zn finger protein Kruppel **(Figure 7.2).** Kruppel is well characterised Zn C2H2 transcription factor with known binding site in *Drosophila* genome. Repeating the SELEX with full length, N, C-term of Topi protein, and other components of tMAC and Kruppel protein can be a good positive control.

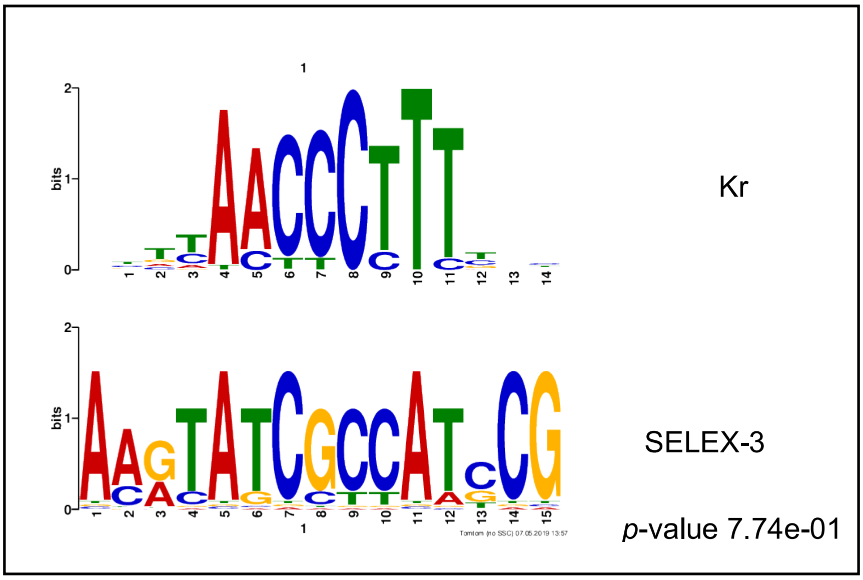


Figure 7.2:- Motif comparison between Kr and Topi C-term SELEX-3. TOMTOM assigns p-value 7.74e-01 for this alignment.

In silico analysis showed that predicted Topi binding site was observed in promoter regions of gene_090 and _074. A motif-3 characterisation results with Topi C-term showed affinity towards both positive as well as negative probes, thus suggests Topi C-term has quite strong affinity for given set of DNA sequences.

Our hypothesis is that tMAC has multiple loose binding sites in promoter region of target gene and it could be possible because of the presence of four different protein with predicted DNA binding domains. The combined effect of DNA binding ability of these four protein contributes and gives the activity, so each individual subunit does not need to bind tightly in promoter of target gene (or low affinity binding sites). This hypothesis could be tested by performing the SELEX using all the proteins with predicted DNA binding domain. One could also use EMSA to validate the discovered motifs for the tMAC subunits. We can also perform the EMSA using non-labelled probes and resolve the protein-DNA complex in SDS-PAGE gel to get more resolution of complex. Repeating the EMSA using labelled probes against Topi C-term would enhance the resolution of the analysis and competitive analysis using different positive probes would tell us about strong and weak affinity towards the given probes.

7.4. Concluding remarks

The first part this thesis has addressed the molecular mechanisms underlying evolution of testis-specific expression of *de novo* genes. I have identified four distinct outcomes of gene expression regulation of gene_090. Firstly a 7 bp deletion in the promoter region of gene_090 is necessary and sufficient to drive the expression in testis; Secondly flanking sites of the 7bp deletion also important for the expression of gene_090; Thirdly adjoining nucleotide of the flanking site creates the binding site for regulatory proteins; fourthly the sequence between the flanking sites is also important for expression of gene_090. The proposed model_2 is possible mode of gene expression of gene_090. For gene_074, no SNPs or indels were identified to a play role in its regulation.

The second part of this thesis has expressed tMAC subunits, identified and characterised the binding site for Topi C-terminal end. I have identified the 15bp *de novo* motif for C-terminal end of Topi protein. The occurrence of the discovered motif-3 in the promoter regions of the *aly* dependent (and independent gene_074) genes and electrophoretic mobility shift analysis of designed oligos using Topi C-term suggest it is genuine.

References.

Adryan, Boris, and Sarah A. Teichmann. 2006. "FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster." *Bioinformatics* 22 (12):1532-1533. doi: 10.1093/bioinformatics/btl143.

Alphey, Luke, Juan Jimenez, Helen White-Cooper, Iain Dawson, Paul Nurse, and David M. Glover. 1992. "twine, a cdc25 homolog that functions in the male and female germline of drosophila." *Cell* 69 (6):977-988. doi: https://doi.org/10.1016/0092-8674(92)90616-K.

Amoutzias, G. D., A. S. Veron, J. Weiner, III, M. Robinson-Rechavi, E. Bornberg-Bauer, S. G. Oliver, and D. L. Robertson. 2006. "One Billion Years of bZIP Transcription Factor Evolution: Conservation and Change in Dimerization and DNA-Binding Site Specificity." *Molecular Biology and Evolution* 24 (3):827-835. doi: 10.1093/molbev/msl211.

Andrews, J., G. G. Bouffard, C. Cheadle, J. Lü, K. G. Becker, and B. Oliver. 2000. "Gene discovery using computational and microarray analysis of transcription in the Drosophila melanogaster testis." *Genome research.* 10 (12):2030-2043.

Arama, Eli, Julie Agapite, and Hermann Steller. 2003. "Caspase Activity and a Specific Cytochrome C Are Required for Sperm Differentiation in Drosophila." *Developmental Cell* 4 (5):687-697. doi: https://doi.org/10.1016/S1534-5807(03)00120-5.

Awe, Stephan, and Renate Renkawitz-Pohl. 2010. "Histone H4 Acetylation is Essential to Proceed from a Histone- to a Protamine-based Chromatin Structure in Spermatid Nuclei of Drosophila melanogaster." *Systems Biology in Reproductive Medicine* 56 (1):44-61. doi: 10.3109/19396360903490790.

Ayyar, Savita, Jianqiao Jiang, Anna Collu, Helen White-Cooper, and Robert A. H. White. 2003. "Drosophila TGIF is essential for developmentally regulated transcription in spermatogenesis." *Development.* 130 (13):2841-2852.

Bailey, Timothy L., James Johnson, Charles E. Grant, and William S. Noble. 2015. "The MEME Suite." *Nucleic acids research* 43 (W1):W39-W49. doi: 10.1093/nar/gkv416.

Banerji, Julian, Sandro Rusconi, and Walter Schaffner. 1981. "Expression of a β-globin gene is enhanced by remote SV40 DNA sequences." *Cell* 27 (2, Part 1):299-308. doi: https://doi.org/10.1016/0092-8674(81)90413-X.

Barreau, Carine, Elizabeth Benson, Elin Gudmannsdottir, Fay Newton, and Helen White-Cooper. 2008. "Post-meiotic transcription in Drosophila testes." *Development* 135 (11):1897-1902.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. "High-Resolution Profiling of Histone Methylations in the Human Genome." *Cell* 129 (4):823-837. doi: https://doi.org/10.1016/j.cell.2007.05.009.

Bauer, Maximilian, and Ralf Metzler. 2012. "Generalized facilitated diffusion model for DNA-binding proteins with search and recognition states." *Biophysical journal* 102 (10):2321-2330. doi: 10.1016/j.bpj.2012.04.008.

Beall, Eileen L., Peter W. Lewis, Maren Bell, Michael Rocha, D. Leanne Jones, and Michael R. Botchan. 2007. "Discovery of tMAC: a Drosophila testis-specific meiotic arrest complex paralogous to Myb-Muv B." *Genes &amp; development.* 21 (8):904-919. doi: 10.1101/gad.1516607.

Beall, Eileen L., J. Robert Manak, Sharleen Zhou, Maren Bell, Joseph S. Lipsick, and Michael R. Botchan. 2002. "Role for a Drosophila Myb-containing protein complex in site-specific DNA replication." *Nature.* 420 (6917):833-837. doi: 10.1038/nature01228.

Begun, David J., Heather A. Lindfors, Andrew D. Kern, and Corbin D. Jones. 2007. "Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade." *Genetics.* 176 (2):1131-1137. doi: 10.1534/genetics.106.069245.

Begun, David J., Heather A. Lindfors, Melissa E. Thompson, and Alisha K. Holloway. 2006. "Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags." *Genetics.* 172 (3):1675-1681. doi: 10.1534/genetics.105.050336.

Betrán, Esther, Kevin Thornton, and Manyuan Long. 2002. "Retroposed new genes out of the X in Drosophila." *Genome research* 12 (12):1854-1859. doi: 10.1101/gr.6049.

Biterge, Burcu, and Robert Schneider. 2014. "Histone variants: key players of chromatin." *Cell and Tissue Research* 356 (3):457-466. doi: 10.1007/s00441-014-1862-4.

Blümer, Nicole, Kay Schreiter, Leonie Hempel, Ansgar Santel, Martin Hollmann, Mireille A. Schäfer, and Renate Renkawitz-Pohl. 2002. "A new translational repression element and unusual transcriptional control regulate expression of don juan during Drosophila spermatogenesis." *Mechanisms of Development* 110 (1):97-112. doi: https://doi.org/10.1016/S0925-4773(01)00577-9.

Bohmann, D., T. J. Bos, A. Admon, T. Nishimura, P. K. Vogt, and R. Tjian. 1987. "Human proto-oncogene c-jun encodes a DNA binding protein with structural and functional properties of transcription factor AP-1." *Science* 238 (4832):1386. doi: 10.1126/science.2825349.

Bonev, Boyan, and Giacomo Cavalli. 2016. "Organization and function of the 3D genome." *Nature Reviews Genetics* 17 (11):661-678. doi: 10.1038/nrg.2016.112.

Boube, Muriel, Bruno Hudry, Clément Immarigeon, Yannick Carrier, Sandra Bernat-Fabre, Samir Merabet, Yacine Graba, Henri-Marc Bourbon, and David L. Cribbs. 2014. "Drosophila melanogaster Hox transcription factors access the RNA polymerase II machinery through direct homeodomain binding to a conserved motif of mediator subunit Med19." *PLoS genetics* 10 (5):e1004303-e1004303. doi: 10.1371/journal.pgen.1004303.

Braun, Pascal, and Josh LaBaer. 2003. "High throughput protein production for functional proteomics." *Trends in Biotechnology* 21 (9):383-388. doi: https://doi.org/10.1016/S0167-7799(03)00189-6.

Busby, Steve, and Richard H. Ebright. 1999. "Transcription activation by catabolite activator protein (CAP)." *Journal of Molecular Biology* 293 (2):199-213. doi: https://doi.org/10.1006/jmbi.1999.3161.

Cai, Jing, Ruoping Zhao, Huifeng Jiang, and Wen Wang. 2008. "De novo origination of a new protein-coding gene in Saccharomyces cerevisiae." *Genetics.* 179 (1):487-496. doi: 10.1534/genetics.107.084491.

Caporilli, Simona, Yachuan Yu, Jianqiao Jiang, and Helen White-Cooper. 2013. "The RNA export factor, Nxt1, is required for tissue specific transcriptional regulation." *PLoS genetics* 9 (6):e1003526-e1003526. doi: 10.1371/journal.pgen.1003526.

Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael A. Calderwood, Muhammed A. Yildirim, Nicolas Simonis, Benoit Charloteaux, César A. Hidalgo, Justin Barbette, Balaji Santhanam, Gloria A. Brar, Jonathan S. Weissman, Aviv Regev, Nicolas Thierry-Mieg, Michael E. Cusick, and Marc Vidal. 2012. "Proto-genes and de novo gene birth." *Nature.* 487 (7407):370-374. doi: 10.1038/nature11184.

Charoensawan, Varodom, Derek Wilson, and Sarah A. Teichmann. 2010. "Lineage-specific expansion of DNA-binding transcription factor families." *Trends in genetics : TIG* 26 (9):388-393. doi: 10.1016/j.tig.2010.06.004.

Chen, Xin, Mark Hiller, Yasemin Sancak, and Margaret T. Fuller. 2005. "Tissue-specific TAFs counteract Polycomb to turn on terminal differentiation." *Science.* 310 (5749):869-872. doi: 10.1126/science.1118101.

Chintapalli, Venkateswara R., Jing Wang, and Julian A. T. Dow. 2007. "Using FlyAtlas to identify better Drosophila melanogaster models of human disease." *Nature genetics.* 39 (6):715-720. doi: 10.1038/ng2049info:doi/10.1038/ng2049.

Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nature Genetics. 2002 Mar;30(3):315–320. pmid:11919562

Courtot, C., C. Fankhauser, V. Simanis, and C. F. Lehner. 1992. "The Drosophila cdc25 homolog twine is required for meiosis." *Development* 116 (2):405.

Demarco, Rafael S., Åsmund H. Eikenes, Kaisa Haglund, and D. Leanne Jones. 2014. "Investigating spermatogenesis in Drosophila melanogaster." *Methods.* 68 (1):218-227. doi: 10.1016/j.ymeth.2014.04.020.

Denslow, S. A., and P. A. Wade. 2007. "The human Mi-2/NuRD complex and gene regulation." *Oncogene* 26 (37):5433-5438. doi: 10.1038/sj.onc.1210611.

Doggett, Karen, Jianqiao Jiang, Gajender Aleti, and Helen White-Cooper. 2011. "Wake-up-call, a lin-52 paralogue, and Always early, a lin-9 homologue physically interact, but have opposing functions in regulating testis-specific gene expression." *Developmental biology.* 355 (2):381-393. doi: 10.1016/j.ydbio.2011.04.030.

Donoghue, Mark T. A., Channa Keshavaiah, Sandesh H. Swamidatta, and Charles Spillane. 2011. "Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana." *BMC evolutionary biology.* 11 (1):47. doi: 10.1186/1471-2148-11-47.

Dror, Iris, Remo Rohs, and Yael Mandel-Gutfreund. 2016. "How motif environment influences transcription factor search dynamics: Finding a needle in a haystack." *BioEssays : news and reviews in molecular, cellular and developmental biology* 38 (7):605-612. doi: 10.1002/bies.201600005.

Dubruille, Raphaëlle, Guillermo A. Orsi, Lætitia Delabaere, Elisabeth Cortier, Pierre Couble, Gabriel A. B. Marais, and Benjamin Loppin. 2010. "Specialization of a Drosophila Capping Protein Essential for the Protection of Sperm Telomeres." *Current Biology* 20 (23):2090-2099. doi: 10.1016/j.cub.2010.11.013.

Eberhart, Charles G., Jean Z. Maines, and Steven A. Wasserman. 1996. "Meiotic cell cycle requirement for a fly homologue of human Deleted in Azoospermia." *Nature* 381 (6585):783-785. doi: 10.1038/381783a0.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, Lorna J. Richardson, Gustavo A. Salazar, Alfredo Smart, Erik L. L. Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C. E. Tosatto, and Robert D. Finn. 2019. "The Pfam protein families database in 2019." *Nucleic acids research* 47 (D1):D427-D432. doi: 10.1093/nar/gky995.

El-Sharnouby, Sherif, Juliet Redhouse, and Robert A. H. White. 2013. "Genome-wide and cell-specific epigenetic analysis challenges the role of polycomb in Drosophila spermatogenesis." *PLoS genetics* 9 (10):e1003842-e1003842. doi: 10.1371/journal.pgen.1003842.

Fabian, Lacramioara, and Julie A. Brill. 2012. "Drosophila spermiogenesis: Big things come from little packages." *Spermatogenesis* 2 (3):197-212. doi: 10.4161/spmg.21798.

Fabrizio, J. J., G. Hime, S. K. Lemmon, and C. Bazinet. 1998. "Genetic dissection of sperm individualization in Drosophila melanogaster." *Development* 125 (10):1833.

Farkas, Rebecca M., Maria Grazia Giansanti, Maurizio Gatti, and Margaret T. Fuller. 2003. "The Drosophila Cog5 homologue is required for cytokinesis, cell elongation, and assembly of specialized Golgi architecture during spermatogenesis." *Molecular biology of the cell* 14 (1):190-200. doi: 10.1091/mbc.e02-06-0343.

Fornes, Oriol, Jaime A. Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A. Richmond, Bhavi P. Modi, Solenne Correard, Marius Gheorghe, Damir Baranašić, Walter Santana-Garcia, Ge Tan, Jeanne Chèneby, Benoit Ballester, François Parcy, Albin Sandelin, Boris Lenhard, Wyeth W. Wasserman, and Anthony Mathelier. 2019. "JASPAR 2020: update of the open-access database of transcription factor binding profiles." *Nucleic Acids Research*. doi: 10.1093/nar/gkz1001.

Gallo, Steven M., Dave T. Gerrard, David Miner, Michael Simich, Benjamin Des Soye, Casey M. Bergman, and Marc S. Halfon. 2011. "REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila." *Nucleic acids research* 39 (Database issue):D118-D123. doi: 10.1093/nar/gkq999.

Gambetta, Maria Cristina, and Eileen E. M. Furlong. 2018. "The Insulator Protein CTCF Is Required for Correct Hox Gene Expression, but Not for Embryonic Development in Drosophila." *Genetics* 210 (1):129. doi: 10.1534/genetics.118.301350.

Gan, Qiang, Iouri Chepelev, Gang Wei, Lama Tarayrah, Kairong Cui, Keji Zhao, and Xin Chen. 2010. "Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq." *Cell research* 20 (7):763-783. doi: 10.1038/cr.2010.64.

Gao, Guanjun, Yan Cheng, Natalia Wesolowska, and Yikang S. Rong. 2011. "Paternal imprint essential for the inheritance of telomere identity in Drosophila." *Proceedings of the National Academy of Sciences* 108 (12):4932. doi: 10.1073/pnas.1016792108.

Gerasimova, Tatiana I., Elissa P. Lei, Ashley M. Bushey, and Victor G. Corces. 2007. "Coordinated Control of dCTCF and gypsy Chromatin Insulators in Drosophila." *Molecular Cell* 28 (5):761-772. doi: 10.1016/j.molcel.2007.09.024.

Gibson, Daniel G., Lei Young, Ray-Yuan Chuang, J. Craig Venter, Clyde A. Hutchison, and Hamilton O. Smith. 2009. "Enzymatic assembly of DNA molecules up to several hundred kilobases." *Nature Methods* 6 (5):343-345. doi: 10.1038/nmeth.1318.

Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. "FIMO: scanning for occurrences of a given motif." *Bioinformatics (Oxford, England)* 27 (7):1017-1018. doi: 10.1093/bioinformatics/btr064.

Gubala, Anna M., Jonathan F. Schmitz, Michael J. Kearns, Tery T. Vinh, Erich Bornberg-Bauer, Mariana F. Wolfner, and Geoffrey D. Findlay. 2017. "The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo." *Molecular biology and evolution* 34 (5):1066-1082. doi: 10.1093/molbev/msx057.

Guo, Ya, Quan Xu, Daniele Canzio, Jia Shou, Jinhuan Li, David U Gorkin, Inkyung Jung, Haiyang Wu, Yanan Zhai, Yuanxiao Tang, Yichao Lu, Yonghu Wu, Zhilian Jia, Wei Li, Michael Q Zhang, Bing Ren, Adrian R Krainer, Tom Maniatis, and Qiang Wu. 2015. "CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function." *Cell* 162 (4):900-910. doi: 10.1016/j.cell.2015.07.038.

Hawley, R. Scott. 2002. "Meiosis: How Male Flies Do Meiosis." *Current Biology* 12 (19):R660-R662. doi: https://doi.org/10.1016/S0960-9822(02)01161-2.

He, Yuan, Jie Fang, Dylan J. Taatjes, and Eva Nogales. 2013. "Structural visualization of key steps in human transcription initiation." *Nature* 495 (7442):481-486. doi: 10.1038/nature11991.

Heinen, Tobias J. A. J., Fabian Staubach, Daniela Häming, and Diethard Tautz. 2009. "Emergence of a New Gene from an Intergenic Region." *Current Biology* 19 (18):1527-1531. doi: https://doi.org/10.1016/j.cub.2009.07.049.

Hemmer, Lucas W., and Justin P. Blumenstiel. 2016. "Holding it together: rapid evolution and positive selection in the synaptonemal complex of Drosophila." *BMC Evolutionary Biology* 16 (1):91. doi: 10.1186/s12862-016-0670-8.

Hendrix, David A., Joung-Woo Hong, Julia Zeitlinger, Daniel S. Rokhsar, and Michael S. Levine. 2008. "Promoter elements associated with RNA Pol II stalling in the Drosophila embryo." *Proceedings of the National Academy of Sciences of the United States of America* 105 (22):7762-7767. doi: 10.1073/pnas.0802406105.

Hillen, Hauke S., Yaroslav I. Morozov, Azadeh Sarfallah, Dmitry Temiakov, and Patrick Cramer. 2017. "Structural Basis of Mitochondrial Transcription Initiation." *Cell* 171 (5):1072-1081.e10. doi: 10.1016/j.cell.2017.10.036.

Hiller, M. A., T. Y. Lin, C. Wood, and M. T. Fuller. 2001. "Developmental regulation of transcription by a tissue-specific TAF homolog." *Genes &amp; development.* 15 (8):1021-1030. doi: 10.1101/gad.869101.

Hiller, Mark, Xin Chen, M. Jodeane Pringle, Martin Suchorolski, Yasemin Sancak, Sridhar Viswanathan, Benjamin Bolival, Ting-Yi Lin, Susan Marino, and Margaret T. Fuller. 2004. "Testis-specific TAF homologs collaborate to control a tissue-specific transcription program." *Development.* 131 (21):5297-5308. doi: 10.1242/dev.01314.

Huang, Pengxiang, Vikas Chandra, and Fraydoon Rastinejad. 2010. "Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics." *Annual review of physiology* 72:247-272. doi: 10.1146/annurev-physiol-021909-135917.

Hughes, Austin L. 2005. "Gene duplication and the origin of novel proteins." *Proceedings of the National Academy of Sciences of the United States of America* 102 (25):8791-8792. doi: 10.1073/pnas.0503922102.

Illingworth, Robert S., Ulrike Gruenewald-Schneider, Shaun Webb, Alastair R. W. Kerr, Keith D. James, Daniel J. Turner, Colin Smith, David J. Harrison, Robert Andrews, and Adrian P. Bird. 2010. "Orphan CpG islands identify numerous conserved promoters in the mammalian genome." *PLoS genetics* 6 (9):e1001134-e1001134. doi: 10.1371/journal.pgen.1001134.

Jacob, François, and Jacques Monod. 1961. "Genetic regulatory mechanisms in the synthesis of proteins." *Journal of Molecular Biology* 3 (3):318-356. doi: https://doi.org/10.1016/S0022-2836(61)80072-7.

Jayaramaiah Raja, Sunil, and Renate Renkawitz-Pohl. 2005. "Replacement by Drosophila melanogaster protamines and Mst77F of histones during chromatin condensation in late spermatids and role of sesame in the removal of these proteins from the male pronucleus." *Molecular and cellular biology* 25 (14):6165-6177. doi: 10.1128/MCB.25.14.6165-6177.2005.

Jiang, Jianqiao, Elizabeth Benson, Nina Bausek, Karen Doggett, and Helen White-Cooper. 2007. "Tombola, a tesmin/TSO1-family protein, regulates transcriptional activation in the Drosophila male germline and physically interacts with always early." *Development.* 134 (8):1549-1559. doi: 10.1242/dev.000521.

Jiang, Jianqiao, and Helen White-Cooper. 2003. "Transcriptional activation in Drosophila spermatogenesis involves the mutually dependent function of aly and a novel meiotic arrest gene cookie monster." *Development.* 130 (3):563-573.

Jiang, Mei, Zhengliang Gao, Jian Wang, and Dmitry I. Nurminsky. 2018. "Evidence for a hierarchical transcriptional circuit in Drosophila male germline involving testis-specific TAF and two gene-specific transcription factors, Mod and Acj6." *FEBS letters* 592 (1):46-59. doi: 10.1002/1873-3468.12937.

Kaessmann, Henrik. 2010. "Origins, evolution, and phenotypic impact of new genes." *Genome research.* 20 (10):1313-1326. doi: 10.1101/gr.101386.109.

Kanippayoor, Rachelle L., Joshua H. M. Alpern, and Amanda J. Moehring. 2013. "Protamines and spermatogenesis in Drosophila and Homo sapiens : A comparative analysis." *Spermatogenesis* 3 (2):e24376-e24376. doi: 10.4161/spmg.24376.

Katzenberger, Rebeccah J., Elizabeth A. Rach, Ashley K. Anderson, Uwe Ohler, and David A. Wassarman. 2012. "The Drosophila Translational Control Element (TCE) is required for high-level transcription of many genes that are specifically expressed in testes." *PloS one* 7 (9):e45009-e45009. doi: 10.1371/journal.pone.0045009.

Kawase, Eihachiro, Marco D. Wong, Bee C. Ding, and Ting Xie. 2004. "Gbb/Bmp signaling is essential for maintaining germline stem cells and for repressing bam transcription in the Drosophila testis." *Development.* 131 (6):1365-1375. doi: 10.1242/dev.01025.

Khan, Aziz, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A. Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R. Kulkarni, Ge Tan, Damir Baranasic, David J. Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier. 2018. "JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework." *Nucleic acids research* 46 (D1):D260-D266. doi: 10.1093/nar/gkx1126.

Kiger, A. A., D. L. Jones, C. Schulz, M. B. Rogers, and M. T. Fuller. 2001. "Stem cell self-renewal specified by JAK-STAT activation in response to a support cell cue." *Science.* 294 (5551):2542-2545. doi: 10.1126/science.1066707.

Kim, Jongmin, Chenggang Lu, Shrividhya Srinivasan, Stephan Awe, Alexander Brehm, and Margaret T. Fuller. 2017. "Blocking promiscuous activation at cryptic promoters directs cell type-specific gene expression." *Science (New York, N.Y.)* 356 (6339):717-721. doi: 10.1126/science.aal3096.

Knowles, David G., and Aoife McLysaght. 2009. "Recent de novo origin of human protein-coding genes." *Genome research.* 19 (10):1752-1759. doi: 10.1101/gr.095026.109.

Korenjak, Michael, Barbie Taylor-Harding, Ulrich K. Binné, John S. Satterlee, Olivier Stevaux, Rein Aasland, Helen White-Cooper, Nick Dyson, and Alexander Brehm. 2004. "Native E2F/RBF Complexes Contain Myb-Interacting Proteins and Repress Transcription of Developmentally Controlled E2F Target Genes." *Cell.* 119 (2):181-193. doi: 10.1016/j.cell.2004.09.034.

Kuhn, Rainer, Claudia Kuhn, Dagmar Börsch, Karl Heinz Glätzer, Ulrich Schäfer, and Mireille Schäfer. 1991. "A cluster of four genes selectively expressed in the male germ line of Drosophila melanogaster." *Mechanisms of development.* 35 (2):143-151. doi: 10.1016/0925-4773(91)90064-D.

Kunert, Natascha, Eugenia Wagner, Magdalena Murawska, Henrike Klinker, Elisabeth Kremmer, and Alexander Brehm. 2009. "dMec: a novel Mi-2 chromatin remodelling complex involved in transcriptional repression." *The EMBO journal* 28 (5):533-544. doi: 10.1038/emboj.2009.3.

Kurshakova, M. M. Auid-Orcid http orcid org, E. N. Nabirochkina, S. G. Georgieva, and D. V. Kopytova. "TRF4, the novel TBP-related protein of Drosophila melanogaster, is concentrated at the endoplasmic reticulum and copurifies with proteins participating in the processes associated with endoplasmic reticulum. LID - 10.1002/jcb.28070 [doi]." (1097-4644 (Electronic)).

Laktionov, P. P., H. White-Cooper, D. A. Maksimov, and S. N. Belyakin. 2014. "Transcription factor Comr acts as a direct activator in the genetic program controlling spermatogenesis in D. melanogaster." *Molecular Biology* 48 (1):130-140. doi: 10.1134/S0026893314010087.

Laktionov, Petr P., Daniil A. Maksimov, Stanislav E. Romanov, Polina A. Antoshina, Olga V. Posukh, Helen White-Cooper, Dmitry E. Koryakov, and Stepan N. Belyakin. 2018. "Genome-wide analysis of gene regulation mechanisms during Drosophila spermatogenesis." *Epigenetics & Chromatin* 11 (1):14. doi: 10.1186/s13072-018-0183-3.

Lawrence, Moyra, Sylvain Daujat, and Robert Schneider. 2016. "Lateral Thinking: How Histone Modifications Regulate Gene Expression." *Trends in Genetics* 32 (1):42-56. doi: https://doi.org/10.1016/j.tig.2015.10.007.

Lee, David J., Stephen D. Minchin, and Stephen J. W. Busby. 2012. "Activating Transcription in Bacteria." *Annual Review of Microbiology* 66 (1):125-152. doi: 10.1146/annurev-micro-092611-150012.

Lenhard, Boris, Albin Sandelin, and Piero Carninci. 2012. "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." *Nature Reviews Genetics* 13 (4):233-245. doi: 10.1038/nrg3163.

Levine, Mia T., Corbin D. Jones, Andrew D. Kern, Heather A. Lindfors, and David J. Begun. 2006. "Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression." *Proceedings of the National Academy of Sciences of the United States of America.* 103 (26):9935-9939. doi: 10.1073/pnas.0509809103.

Lewis, Peter W., Eileen L. Beall, Tracey C. Fleischer, Daphne Georlette, Andrew J. Link, and Michael R. Botchan. 2004. "Identification of a Drosophila Myb-E2F2/RBF transcriptional repressor complex." *Genes &amp; development.* 18 (23):2929-2940. doi: 10.1101/gad.1255204.

Li, Victor C., Jerel C. Davis, Kapa Lenkov, Benjamin Bolival, Margaret T. Fuller, and Dmitri A. Petrov. 2009. "Molecular evolution of the testis TAFs of Drosophila." *Molecular biology and evolution* 26 (5):1103-1116. doi: 10.1093/molbev/msp030.

Li, Xiao-yong, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A. Pollard, Venky N. Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L. Luengo Hendriks, Hou Cheng Chu, Nobuo Ogawa, William Inwood, Victor Sementchenko, Amy Beaton, Richard Weiszmann, Susan E. Celniker, David W. Knowles, Tom Gingeras, Terence P. Speed, Michael B. Eisen, and Mark D. Biggin. 2008. "Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm." *PLoS biology* 6 (2):e27-e27. doi: 10.1371/journal.pbio.0060027.

Lin, T. Y., S. Viswanathan, C. Wood, P. G. Wilson, N. Wolf, and M. T. Fuller. 1996. "Coordinate developmental control of the meiotic cell cycle and spermatid differentiation in Drosophila males." *Development.* 122 (4):1331-1341.

Long, Manyuan, Nicholas W. VanKuren, Sidi Chen, and Maria D. Vibranovski. 2013. "New gene evolution: little did we know." *Annual review of genetics* 47:307-333. doi: 10.1146/annurev-genet-111212-133301.

Louder, Robert K., Yuan He, José Ramón López-Blanco, Jie Fang, Pablo Chacón, and Eva Nogales. 2016. "Structure of promoter-bound TFIID and model of human pre-initiation complex assembly." *Nature* 531 (7596):604-609. doi: 10.1038/nature17394.

Lu, Chenggang, Jongmin Kim, and Margaret T. Fuller. 2013. "The polyubiquitin gene Ubi-p63E is essential for male meiotic cell cycle progression and germ cell differentiation in Drosophila." *Development (Cambridge, England)* 140 (17):3522-3531. doi: 10.1242/dev.098947.

Ma, Jun. 2011. "Transcriptional activators and activation mechanisms." *Protein & cell* 2 (11):879-888. doi: 10.1007/s13238-011-1101-7.

Maines, Jean Z., and Steven A. Wasserman. 1999. "Post-transcriptional regulation of the meiotic Cdc25 protein Twine by the Dazl orthologue Boule." *Nature Cell Biology* 1 (3):171-174. doi: 10.1038/11091.

Mathelier, Anthony, Wenqiang Shi, and Wyeth W. Wasserman. 2015. "Identification of altered cis-regulatory elements in human disease." *Trends in Genetics* 31 (2):67-76. doi: https://doi.org/10.1016/j.tig.2014.12.003.

Metcalf, Chad E., and David A. Wassarman. 2007. "Nucleolar colocalization of TAF1 and testis-specific TAFs during Drosophila spermatogenesis." *Developmental dynamics.* 236 (10):2836-2843. doi: 10.1002/dvdy.21294.

Michiels, Frits, Alexander Gasch, Barbara Kaltschmidt, and Renate Renkawitz-Pohl. 1989. "A 14 bp promoter element directs the testis specificity of the Drosophila beta 2 tubulin gene." *The EMBO journal* 8 (5):1559.

Moon, Sungjin, Bongki Cho, Su-Hong Min, Daekee Lee, and Yun Doo Chung. 2011. "The THO complex is required for nucleolar integrity in Drosophila spermatocytes." *Development* 138 (17):3835. doi: 10.1242/dev.056945.

Morgunova, Ekaterina, Yimeng Yin, Arttu Jolma, Kashyap Dave, Bernhard Schmierer, Alexander Popov, Nadejda Eremina, Lennart Nilsson, and Jussi Taipale. 2015. "Structural insights into the DNA-binding specificity of E2F family transcription factors." *Nature communications* 6:10050-10050. doi: 10.1038/ncomms10050.

Murawska, Magdalena, Natascha Kunert, Joke van Vugt, Gernot Längst, Elisabeth Kremmer, Colin Logie, and Alexander Brehm. 2008. "dCHD3, a Novel ATP-Dependent Chromatin Remodeler Associated with Sites of Active Transcription." *Molecular and Cellular Biology* 28 (8):2745. doi: 10.1128/MCB.01839-07.

Murphy, M. B., S. T. Fuller, P. M. Richardson, and S. A. Doyle. 2003. "An improved method for the in vitro evolution of aptamers and applications in protein detection and purification." *Nucleic Acids Res* 31 (18):e110. doi: 10.1093/nar/gng110.

Narendra, Varun, Pedro P. Rocha, Disi An, Ramya Raviram, Jane A. Skok, Esteban O. Mazzoni, and Danny Reinberg. 2015. "CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation." *Science* 347 (6225):1017. doi: 10.1126/science.1262088.

Neme, Rafik, and Diethard Tautz. 2013. "Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution." *BMC genomics.* 14 (1):117. doi: 10.1186/1471-2164-14-117.

Nikolov, D. B, and S. K Burley. 1997. "RNA polymerase II transcription initiation: A structural view." *Proceedings of the National Academy of Sciences* 94 (1):15. doi: 10.1073/pnas.94.1.15.

Noguchi, Tatsuhiko, and Kathryn G. Miller. 2003. "A role for actin dynamics in individualization during spermatogenesis in Drosophila melanogaster." *Development* 130 (9):1805. doi: 10.1242/dev.00406.

Noyes, Marcus B., Xiangdong Meng, Atsuya Wakabayashi, Saurabh Sinha, Michael H. Brodsky, and Scot A. Wolfe. 2008. "A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system." *Nucleic acids research* 36 (8):2547-2560. doi: 10.1093/nar/gkn048.

Nurminsky, Dmitry I., Maria V. Nurminskaya, Daniel De Aguiar, and Daniel L. Hartl. 1998. "Selective sweep of a newly evolved sperm-specific gene in Drosophila." *Nature* 396 (6711):572-575. doi: 10.1038/25126.

Ohlstein, B., C. A. Lavoie, O. Vef, E. Gateff, and D. M. McKearin. 2000. "The Drosophila cystoblast differentiation factor, benign gonial cell neoplasm, is related to DExH-box proteins and interacts genetically with bag-of-marbles." *Genetics* 155 (4):1809-1819.

Oliphant, A. R., C. J. Brandl, and K. Struhl. 1989. "Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein." *Molecular and cellular biology* 9 (7):2944-2949. doi: 10.1128/mcb.9.7.2944.

Olivieri, G., and A. Olivieri. 1965. "Autoradiographic study of nucleic acid synthesis during spermatogenesis in Drosophila melanogaster." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 2 (4):366-380. doi: https://doi.org/10.1016/0027-5107(65)90072-2.

Ozer, Abdullah, John M. Pagano, and John T. Lis. 2014. "New Technologies Provide Quantum Changes in the Scale, Speed, and Success of SELEX Methods and Aptamer Characterization." *Molecular therapy. Nucleic acids* 3 (8):e183-e183. doi: 10.1038/mtna.2014.34.

Palmieri, Nicola, Carolin Kosiol, and Christian Schlötterer. 2014. "The life cycle of Drosophila orphan genes." *eLife* 3:e01311-e01311. doi: 10.7554/eLife.01311.

Perezgasga, Lucia, JianQiao Jiang, Benjamin Bolival, Mark Hiller, Elizabeth Benson, Margaret T. Fuller, and Helen White-Cooper. 2004. "Regulation of transcription of meiotic cell cycle and terminal differentiation genes by the testis-specific Zn-finger protein matotopetli." *Development* 131 (8):1691. doi: 10.1242/dev.01032.

Persikov, Anton V., Elizabeth F. Rowland, Benjamin L. Oakes, Mona Singh, and Marcus B. Noyes. 2014. "Deep sequencing of large library selections allows computational discovery of diverse sets of zinc fingers that bind common targets." *Nucleic acids research* 42 (3):1497-1508. doi: 10.1093/nar/gkt1034.

Phillips, D. M. 1970. "Insect sperm: their structure and morphogenesis." *The Journal of cell biology* 44 (2):243-277. doi: 10.1083/jcb.44.2.243.

Plass, Christoph, Stefan M. Pfister, Anders M. Lindroth, Olga Bogatyrova, Rainer Claus, and Peter Lichter. 2013. "Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer." *Nature Reviews Genetics* 14 (11):765-780. doi: 10.1038/nrg3554.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji

Yamada, Daniel R. Mende, Junhua Li, Junming Xu, Shaochuan Li, Dongfang Li, Jianjun Cao, Bo Wang, Huiqing Liang, Huisong Zheng, Yinlong Xie, Julien Tap, Patricia Lepage, Marcelo Bertalan, Jean-Michel Batto, Torben Hansen, Denis Le Paslier, Allan Linneberg, H. Bjørn Nielsen, Eric Pelletier, Pierre Renault, Thomas Sicheritz-Ponten, Keith Turner, Hongmei Zhu, Chang Yu, Shengting Li, Min Jian, Yan Zhou, Yingrui Li, Xiuqing Zhang, Songgang Li, Nan Qin, Huanming Yang, Jian Wang, Søren Brunak, Joel Doré, Francisco Guarner, Karsten Kristiansen, Oluf Pedersen, Julian Parkhill, Jean Weissenbach, H. I. T. Consortium Meta, Peer Bork, S. Dusko Ehrlich, and Jun Wang. 2010. "A human gut microbial gene catalogue established by metagenomic sequencing." *Nature* 464 (7285):59-65. doi: 10.1038/nature08821.

Ranz, José María, Ana Rita Ponce, Daniel L. Hartl, and Dmitry Nurminsky. 2003. "Origin and evolution of a new gene expressed in the Drosophila sperm axoneme." In *Origin and Evolution of New Gene Functions*, 233-244. Springer.

Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez Lieberman Aiden. 2014. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell* 159 (7):1665-1680. doi: 10.1016/j.cell.2014.11.021.

Rathke, Christina, Willy M. Baarends, Sunil Jayaramaiah-Raja, Marek Bartkuhn, Rainer Renkawitz, and Renate Renkawitz-Pohl. 2007. "Transition from a nucleosome-based to a protamine-based chromatin configuration during spermiogenesis in Drosophila." *Journal of Cell Science* 120 (9):1689. doi: 10.1242/jcs.004663.

Reinhardt, Josephine A., Betty M. Wanjiru, Alicia T. Brant, Perot Saelao, David J. Begun, Corbin D. Jones, and Esther Betran. 2013. "De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences." *PLOS genetics.* 9 (10):e1003860. doi: 10.1371/journal.pgen.1003860.

Rossetto, Dorine, Nikita Avvakumov, and Jacques Côté. 2012. "Histone phosphorylation: a chromatin modification involved in diverse nuclear events." *Epigenetics* 7 (10):1098-1108. doi: 10.4161/epi.21975.

Ruiz-Orera, Jorge, Jessica Hernandez-Rodriguez, Cristina Chiva, Eduard Sabidó, Ivanela Kondova, Ronald Bontrop, Tomàs Marqués-Bonet, MMar Albà, and James Noonan. 2015. "Origins of De Novo Genes in Human and Chimpanzee." *PLOS genetics.* 11 (12):e1005721. doi: 10.1371/journal.pgen.1005721.

Sabari, Benjamin R., Di Zhang, C. David Allis, and Yingming Zhao. 2017. "Metabolic regulation of gene expression through histone acylations." *Nature reviews. Molecular cell biology* 18 (2):90-101. doi: 10.1038/nrm.2016.140.

Santel, A., J. Kaufmann, R. Hyland, and R. Renkawitz-Pohl. 2000. "The initiator element of the Drosophila beta2 tubulin gene core promoter contributes to gene expression in vivo but is not required for male germ-cell specific expression." *Nucleic acids research* 28 (6):1439-1446. doi: 10.1093/nar/28.6.1439.

Schlötterer, Christian. 2015. "Genes from scratch – the evolutionary fate of de novo genes." *Trends in genetics.* 31 (4):215-219. doi: 10.1016/j.tig.2015.02.007.

Schmidt, Hugo G., Sven Sewitz, Steven S. Andrews, and Karen Lipkow. 2014. "An integrated model of transcription factor diffusion shows the importance of intersegmental transfer and quaternary protein structure for target site finding." *PloS one* 9 (10):e108575-e108575. doi: 10.1371/journal.pone.0108575.

Schöne, Stefanie, Marcel Jurk, Mahdi Bagherpoor Helabad, Iris Dror, Isabelle Lebars, Bruno Kieffer, Petra Imhof, Remo Rohs, Martin Vingron, Morgane Thomas-Chollier, and Sebastiaan H. Meijsing. 2016. "Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity." *Nature communications* 7:12621-12621. doi: 10.1038/ncomms12621.

Schütze, Tatjana, Florian Rubelt, Julia Repkow, Nicole Greiner, Volker A. Erdmann, Hans Lehrach, Zoltán Konthur, and Jörn Glökler. 2011. "A streamlined protocol for emulsion polymerase chain reaction and subsequent purification." *Analytical Biochemistry* 410 (1):155-157. doi: https://doi.org/10.1016/j.ab.2010.11.029.

Shao, Keke, Weifeng Ding, Feng Wang, Haiquan Li, Da Ma, and Huimin Wang. 2011. "Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection." *PloS one* 6 (9):e24910-e24910. doi: 10.1371/journal.pone.0024910.

Sigrist, Christian J. A., Edouard de Castro, Lorenzo Cerutti, Béatrice A. Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. 2013. "New and continuing developments at PROSITE." *Nucleic acids research* 41 (Database issue):D344-D347. doi: 10.1093/nar/gks1067.

Sigrist, S., H. Jacobs, R. Stratmann, and C. F. Lehner. 1995. "Exit from mitosis is regulated by Drosophila fizzy and the sequential destruction of cyclins A, B and B3." *The EMBO journal* 14 (19):4827-4838.

Sinden, D., M. Badgett, J. Fry, T. Jones, R. Palmen, X. Sheng, A. Simmons, E. Matunis, and M. Wawersik. 2012. "Jak-STAT regulation of cyst stem cell development in the Drosophila testis." *Developmental biology* 372 (1):5-16. doi: 10.1016/j.ydbio.2012.09.009.

Singh, Shree Ram, Zhiyu Zheng, Hong Wang, Su-Wan Oh, Xiu Chen, and Steven X. Hou. 2010. "Competitiveness for the niche and mutual dependence of the germline and somatic stem cells in the Drosophila testis are regulated by the JAK/STAT signaling." *Journal of cellular physiology* 223 (2):500-510. doi: 10.1002/jcp.22073.

Slattery, Matthew, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. 2014. "Absence of a simple code: how transcription factors read the genome." *Trends in biochemical sciences* 39 (9):381-399. doi: 10.1016/j.tibs.2014.07.002.

Spradling, Allan, Margaret T. Fuller, Robert E. Braun, and Shosei Yoshida. 2011. "Germline stem cells." *Cold Spring Harbor perspectives in biology* 3 (11):a002642-a002642. doi: 10.1101/cshperspect.a002642.

Stevens, Tim J., David Lando, Srinjan Basu, Liam P. Atkinson, Yang Cao, Steven F. Lee, Martin Leeb, Kai J. Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, Julie Cramard, Andre J. Faure, Meryem Ralser, Enrique Blanco, Lluis Morey, Miriam Sansó, Matthieu G. S. Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, Brian Hendrich, Dave Klenerman, and Ernest D. Laue. 2017. "3D structures of individual mammalian genomes studied by single-cell Hi-C." *Nature* 544 (7648):59-64. doi: 10.1038/nature21429.

Thummel, Carl S., Anne M. Boulet, and Howard D. Lipshitz. 1988. "Vectors for Drosophila P-element-mediated transformation and tissue culture transfection." *Gene* 74 (2):445-456. doi: https://doi.org/10.1016/0378-1119(88)90177-1.

Toll-Riera, M., N. Bosch, N. Bellora, R. Castelo, L. Armengol, X. Estivill, and M. Mar Alba. 2008. "Origin of Primate Orphan Genes: A Comparative Genomics Approach." *Molecular biology and evolution.* 26 (3):603-612. doi: 10.1093/molbev/msn281.

Tsankova, Nadia, William Renthal, Arvind Kumar, and Eric J. Nestler. 2007. "Epigenetic regulation in psychiatric disorders." *Nature Reviews Neuroscience* 8 (5):355-367. doi: 10.1038/nrn2132.

Tuerk, C., and L. Gold. 1990. "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." *Science* 249 (4968):505. doi: 10.1126/science.2200121.

Tulina, N., and E. Matunis. 2001. "Control of stem cell self-renewal in Drosophila spermatogenesis by JAK-STAT signaling." *Science.* 294 (5551):2546-2549. doi: 10.1126/science.1066700.

Vibranovski, Maria D., Hedibert F. Lopes, Timothy L. Karr, and Manyuan Long. 2009. "Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes." *PLoS genetics* 5 (11):e1000731.

Villar, D., Flicek, P. & Odom, D. Evolution of transcription factor binding in metazoans — mechanisms and functional implications. *Nat Rev Genet* **15,** 221–233 (2014). https://doi.org/10.1038/nrg3481

Voog, Justin, Cecilia D'Alterio, and D. Leanne Jones. 2008. "Multipotent somatic stem cells contribute to the stem cell niche in the Drosophila testis." *Nature.* 454 (7208):1132-1136. doi: 10.1038/nature07173info:doi/10.1038/nature07173.

Wagner, James R., Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen, and Mathieu Blanchette. 2014. "The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts." *Genome Biology* 15 (2):R37. doi: 10.1186/gb-2014-15-2-r37.

Wakimoto, Barbara T., Dan L. Lindsley, and Cheryl Herrera. 2004. "Toward a comprehensive genetic analysis of male fertility in Drosophila melanogaster." *Genetics* 167 (1):207-216. doi: 10.1534/genetics.167.1.207.

Wang, Zhaohui, and Richard S. Mann. 2003. "Requirement for two nearly identical TGIF-related homeobox genes in Drosophila spermatogenesis." *Development.* 130 (13):2853-2865.

Wasserman, Wyeth W., and Albin Sandelin. 2004. "Applied bioinformatics for the identification of regulatory elements." *Nature Reviews Genetics* 5 (4):276-287. doi: 10.1038/nrg1315.

Wei, Gong-Hong, Gwenael Badis, Michael F. Berger, Teemu Kivioja, Kimmo Palin, Martin Enge, Martin Bonke, Arttu Jolma, Markku Varjosalo, Andrew R. Gehrke, Jian Yan, Shaheynoor Talukder, Mikko Turunen, Mikko Taipale, Hendrik G. Stunnenberg, Esko Ukkonen, Timothy R. Hughes, Martha L. Bulyk, and Jussi Taipale. 2010. "Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo." *The EMBO journal* 29 (13):2147-2160. doi: 10.1038/emboj.2010.106.

Wei, Jiufeng, Guodong Li, Shuwei Dang, Yuhui Zhou, Kai Zeng, and Ming Liu. 2016. "Discovery and Validation of Hypermethylated Markers for Colorectal Cancer." *Disease markers* 2016:2192853-2192853. doi: 10.1155/2016/2192853.

White-Cooper, H., L. Alphey, and D. M. Glover. 1993. "The cdc25 homologue twine is required for only some aspects of the entry into meiosis in Drosophila." *Journal of Cell Science* 106 (4):1035.

White-Cooper, H., D. Leroy, A. MacQueen, and M. T. Fuller. 2000. "Transcription of meiotic cell cycle and terminal differentiation genes depends on a conserved chromatin associated protein, whose nuclear localisation is regulated." *Development.* 127 (24):5463-5473.

White-Cooper, H., M. A. Schäfer, L. S. Alphey, and M. T. Fuller. 1998. "Transcriptional and post-transcriptional control mechanisms coordinate the onset of spermatid differentiation with meiosis I in Drosophila." *Development.* 125 (1):125-134.

White-Cooper, Helen. 2010. "Molecular mechanisms of gene regulation during Drosophila spermatogenesis." *Reproduction the official journal of the Society for the Study of Fertility.* 139 (1):11-21. doi: 10.1530/REP-09-0083.

White-Cooper, Helen. 2012. "Tissue, cell type and stage-specific ectopic gene expression and RNAi induction in the Drosophila testis." *Spermatogenesis* 2 (1):11-22. doi: 10.4161/spmg.19088.

Williamson, Iain, Lauren Kane, Paul S. Devenney, Ilya M. Flyamer, Eve Anderson, Fiona Kilanowski, Robert E. Hill, Wendy A. Bickmore, and Laura A. Lettice. 2019. "Developmentally regulated Shh expression is robust to TAD perturbations." *Development* 146 (19):dev179523. doi: 10.1242/dev.179523.

Williamson, Iain, Laura A. Lettice, Robert E. Hill, and Wendy A. Bickmore. 2016. "Shh and ZRS enhancer colocalisation is specific to the zone of polarising activity." *Development* 143 (16):2994. doi: 10.1242/dev.139188.

Wingender, E., P. Dietze, H. Karas, and R. Knüppel. 1996. "TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites." *Nucleic Acids Research* 24 (1):238-241. doi: 10.1093/nar/24.1.238.

Witt, Evan, Sigi Benjamin, Nicolas Svetec, and Li Zhao. 2019. "Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila." *eLife* 8:e47138. doi: 10.7554/eLife.47138.

Wj, Guo. 2007. "Significant comparative characteristics between orphan and nonorphan genes in the rice (Oryza sativa L.) genome." *Comparative and functional genomics.*

Xiao, Wenfei, Hongbo Liu, Yu Li, Xianghua Li, Caiguo Xu, Manyuan Long, Shiping Wang, and Hany A. El-Shemy. 2009. "A Rice Gene of De Novo Origin Negatively Regulates Pathogen-Induced Defense Response." *PloS one.* 4 (2):e4603. doi: 10.1371/journal.pone.0004603.

Yang, Jun, Larry Porter, and John Rawls. 1995. "Expression of the dihydroorotate dehydrogenase gene, dhod, during spermatogenesis in Drosophila melanogaster." *Molecular and General Genetics MGG* 246 (3):334-341. doi: 10.1007/BF00288606.

Yona, Avihu H., Eric J. Alm, and Jeff Gore. 2018. "Random sequences rapidly evolve into de novo promoters." *Nature communications* 9 (1):1530-1530. doi: 10.1038/s41467-018-04026-w.

Zenkin, Nikolay, and Yulia Yuzenkova. 2015. "New Insights into the Functions of Transcription Factors that Bind the RNA Polymerase Secondary Channel." *Biomolecules* 5 (3):1195-1209. doi: 10.3390/biom5031195.

Zhang, W., Gao, Y., Long, M. *et al.* Origination and evolution of orphan genes and *de novo* genes in the genome of *Caenorhabditis elegans*. *Sci. China Life Sci.* **62,** 579–593 (2019). https://doi.org/10.1007/s11427-019-9482-0

Zhao, Li, Perot Saelao, Corbin D. Jones, and David J. Begun. 2014. "Origin and spread of de novo genes in Drosophila melanogaster populations." *Science.* 343 (6172):769-772. doi: 10.1126/science.1248286.

Zhao, Shaowei, Di Chen, Qing Geng, and Zhaohui Wang. 2013. "The highly conserved LAMMER/CLK2 protein kinases prevent germ cell overproliferation in Drosophila." *Developmental Biology* 376 (2):163-170. doi: https://doi.org/10.1016/j.ydbio.2013.01.023.

Zhao, Shuai, Xingrun Zhang, and Haitao Li. 2018. "Beyond histone acetylation—writing and erasing histone acylations." *Current Opinion in Structural Biology* 53:169-177. doi: https://doi.org/10.1016/j.sbi.2018.10.001.

Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. Regulatory Variation Within and Between Species. Annual Review of Genomics and Human Genetics. 2011;12(1):327–346. pmid:21721942

Zhou, Qi, Guojie Zhang, Yue Zhang, Shiyu Xu, Ruoping Zhao, Zubing Zhan, Xin Li, Yun Ding, Shuang Yang, and Wen Wang. 2008. "On the origin of new genes in Drosophila." *Genome research.* 18 (9):1446-1455. doi: 10.1101/gr.076588.108.

Zhu, Qianzheng, and Altaf A. Wani. 2010. "Histone modifications: crucial elements for damage response and chromatin restoration." *Journal of cellular physiology* 223 (2):283-288. doi: 10.1002/jcp.22060.

Zubay, G., D. Schwartz, and J. Beckwith. 1970. "Mechanism of activation of catabolite-sensitive genes: a positive control system." *Proceedings of the National Academy of Sciences of the United States of America* 66 (1):104-110. doi: 10.1073/pnas.66.1.104.

Appendices.

Appendix 1- Primers

Table 5:- List of primers

Primer Name	Sequence (5'-3')	Refences
Pst1+attB For	CTGCAGGTCGACGATGTAGGTCACGG	N.A
***Pst1*+attB Rev**	CTGCAGGTCGACATGCCCGCCGTGAC	N.A
***Pst1*+attB For**	CCAAGCTTGCATGCCTGCAGGTCGACGATGTAGGTCACGG	N.A
Pst1+attB Rev	TAACTTGCACTTTACTGCAGGTCGACATGCCCGCCGTGAC	N.A
pCaSpeR b-gal For	ACTTCAAATACCCTTGGATCG	N.A
pCaSpeR b-gal Rev	CTTGTTGGTCAAAGTAAAC	N.A
SELEX For	GGTATTGAGGGTCGCATC	(Murphy et al. 2003)
SELEX Rev	AGAGGAGAGTTAGAGCCATC	(Murphy et al. 2003)
Topi FL NdeI	AGCGGCCTGGTGCCGCGCGGCAGCCATATGAAAGTCAAAGTTTCGGGTGAATATACG	N.A
Topi FL HindIII	TGGTGGTGCTCGAGTGCGGCCGCAAGCTTTACAAATAGGTATCCGAAAATATCGGCTT	N.A
Topi N-term NdeI	AGCGGCCTGGTGCCGCGCGGCAGCCATATGAAAGTCAAAGTTTCGGGTGAATATACGC	N.A
Topi N-term HindIII	TGGTGGTGCTCGAGTGCGGCCGCAAGCT TTAGTTGAGATGCTTCTCCTCGCTGTGCGC	N.A
Topi C-term NdeI	AGCGGCCTGGTGCCGCGCGGCAGCCATATGAAAAAGAAGCGCGAAAAGGAAACGCGC	N.A
Topi C-term HindIII	TGGTGGTGCTCGAGTGCGGCCGCAAGCT TTACAAATAGGTATCCGAAAATATCGGCTT	N.A
Comr FL NdeI	AGCGGCCTGGTGCCGCGCGGCAGCCATATGTCGGGGAACCAAGACACTTTGGGCCAG	N.A
Comr FL HindIII	TGGTGGTGCTCGAGTGCGGCCGCAAGCTTCAACGAGGATTTCGCTTGGAGTTGCGTAG	N.A

Comr DBD NdeI	AGCGGCCTGGTGCCGCGCGGCAGCCATATG CATCCTCTACTCCATTATGCA	N.A
Comr DBD HindIII	TGGTGGTGCTCGAGTGCGGCCGCAAGCT TCATCCGCCGTTCCCGGGCTTGA	N.A
Tomb FL NdeI	AGCGGCCTGGTGCCGCGCGGCAGCCATATGCCATCGCCCAAGAAAAGAAGTGTGGATA	N.A
Tomb FL HindIII	TGGTGGTGCTCGAGTGCGGCCGCAAGCTTTAATAATAATCCTTGGCTGTACTCGCATA	N.A
Tomb DBD NdeI	AGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGCCATCGCCCAAGAAAAGAAGTGTGGAT	N.A
Tomb DBD HindIII	AGCAGCGGCCTGGTGCCGCGCGGCAGCCATGATGCCCGCCTTGGCAGCTTA	N.A
Achi/Vis FL NdeI	AGCGGCCTGGTGCCGCGCGGCAGCCATATGATCTCGCCGGAACAAGAAGAGGTCAACA	N.A
Achi/Vis FL HindIII	TGGTGGTGCTCGAGTGCGGCCGCAAGCTCTAGTCTCCCATGTAAACGAAATCGTCATC	N.A
Achi/Vis DBD NdeI	AGCGGCCTGGTGCCGCGCGGCAGCCATATGTCGGACTCAAGTTTGGACCAGGATTCTCT	N.A
Achi/Vis DBD HindIII	TGGTGGTGCTCGAGTGCGGCCGCAAGCTCTACAGAGGATCGTTGCCCTCGCGCCTGA	N.A
T7 Promoter	CCTATAGTGAGTCGTATTA	N.A
T7 Term	CCGCTGAGCAATAACTAGC	N.A
LacZ 635 F	CCGCGAGGTGCGGATTGAAA	N.A
LacZ 635 T3R	GCAACGAATTAACCCTCACTAAAGGGTACCCATCGCGTGGGCGTAT	N.A
LacZ 670 F	CAACCCGTGGTCGGCTTACG	N.A
LacZ 670 T3R	GCAACGAATTAACCCTCACTAAAGGGCGTTAGGGTCAATGCGGG	N.A
LacZ 788 F	ACCGATCGCCCTTCCCAACA	N.A
LacZ 788 T3R	GCAACGAATTAACCCTCACTAAAGGGATTTCGGCGCTCCACAG	N.A

Appendix 2- Construct design for gene_090 and gene_074

Table 6:- List of synthetic DNA constructs for gene_090 and gene_074

Construct Name	Sequence 5’-3’	Reference
090_A	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATACTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAATCCATCTGAGCCGTGTCTTGCCGC	(Zhao et al. 2014)
090_B	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTCCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAATAATCTTCGCATATAACTAATATATCTTTATCTATATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGACAATCATATGATTTTTTGTGTTTCAGATACATCAAAAATGCATAGTACTACATAATAATGGCTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGCTGAGCTATCTATTATATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACAAGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	(Zhao et al. 2014)
090_C	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATAATAATGGCTTTTGCATTTTGAGTAACATACATAA	N.A

	ATCAAAATGCTAATATGTTGATCTATCTATTACATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	
090_D	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTCCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAATAATCTTCGCATATAACTAATATATCTTTATCTATATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGACAATCATATGATTTTTTGTGTTTCAGATACATCAAAAATGCATAGTACTACATACTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGCTGAGCTATCTATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACAAGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_E	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATAATAATGGCTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_F	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATACTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCTATTACATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_G	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCAT	N.A

	ATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATACTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	
090_H	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTCCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAATAATCTTCGCATATAACTAATATATCTTTATCTATATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGACAATCATATGATTTTTTGTGTTTCAGATACATCAAAAATGCATAGTACTACATACTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGCTGAGCTATCTATTATATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACAAGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_I	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATACTACATACTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_J	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTCCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAATAATCTTCGCATATAACTAATATATCTTTATCTATATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGACAATCATATGATTTTTTGTGTTTCAGATACATCAAAAATGCATAGTACTACATACTACATACTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGCTGAGCTATCTATTATATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACAAGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	N.A

090_K	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATACTTTTGCCTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_L	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTCCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAATAATCTTCGCATATAACTAATATATCTTTATCTATATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGACAATCATATGATTTTTTGTGTTTCAGATACATCAAAAATGCATAGTACTACATACTTTTGCCTTTTGCATTTTGAGTAACATACATAAATAAAAATGCTAATATGCTGAGCTATCTATTATATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACAAGTAGAAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_M	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATATAGAGATCTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_N	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATACCATTATCTTTTGCATTTTGAGTAACATACATAAA	N.A

	TCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAATCCATCTGAGCCGTGTCTTGCCGC	
090_O	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAATAATGGAAAATGCATAGTACTACATACTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_P	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATATAATGGGCATAGTACTACATACTTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAATCCATCTGAGCCGTGTCTTGCCGC	N.A
090_Q	GGGTATTCCGCAGACACTTGCCCAAATGTTCGCCACTTGTGTTTGGGTCCATTGTTTTGGGTTCTTCTTGGCTTCTGTGTAGTGGAAAACTGCATCGGCATCAACCATTCTTTGGGTCCCAATCCATAGATGCACTGGAAGTATTTAGTAATCTTCGCATATAACTAATATATCTTTATCTTTATAGTTTATAGTTCAATACTTGGATTGGATGTTACTTATAAGCTTTATTAAGAAAATCATTTGATTTTTTTTTGTTTCAGATACATTAAAAATGCATAGTACTACATCATTTTGCATTTTGAGTAACATACATAAATCAAAATGCTAATATGTTGATCTATCCATATATGTGTTAAAATTAAGTAAAAGATTTTCTTCGCTGTGTACATGTAGAATCCATCTGAGCCGTGTCTTGCCGC	N.A
074_ref	AGGGCGTGAAATTGTTTAAAGTACAGTAAAGTTTGCGTGCAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGATCGTTGCGTAAAAGTTTAAAACATTCTCTGTCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGTC	N.A

	GCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCCAGCATTTTTTTTTATTTTACC CGACTGGCGGCAAATTACATTTATTTTTTCTTTGTCATAATTGCGGATGTTAAGCCAAAT	
074_517	AGGGCGTGAAATTGTTTAAAGTACAGTAAAGTTTGCGTGGAAAACTTTTTCCAGCACCGCGTTGCGCTGAAAGATCGTT GCGTAAAAGTTTAAAACATTCTCTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCAGTGTGCCGTGCCAAATTGCCAGCATTTTTTTTTTATTTTA CCCGACTGGCGGCAAATTACATTTATTTTTTCTTTGTC ATAATTGCGGATGTTAAGCCAAAT	**(**Zhao et al. 2014)
074_aly1	AGGGCGTGAAATTGTTTAAAGTACAGTAAAGTTTGCGTGGAAAACTTTTTCCAGCACCGCATTGCGCTGAAAGATCGTT GCGTAAAAGTTTAAAACATTCTGTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCATTGTGCCGTGCCAAATTGCCAGCATTTTTTTTTATTTTAC CCGACTGGCGGCAAATTACATTTATTTTTTCTTTGTC ATAATTGCGGATGTTAAGCCAAAT	N.A
074_517A2	AGGGCGTGAAATTGTTTAAAGTACAGTAAAGTTTGCGTGGAAAACTTTTTCCAGCACCGCGTTGCGCTGAAAGATCGTT GCGTAAAAGTTTAAAACATTCTCTGCCCGGTTAGAGCACTCGATTCTTATCGAGCATGGCCAAACGTTTGGCAAATTGT CGCCATAAATAATAAATATTAACAAAAGCCGGCGACATCAGTGTGCCGTGCCAAATTGCCAGCATTTTTTTTTTTATTTT ACCCGACTGGCGGCAAATTACATTTATTTTTTCTTTGTC ATAATTGCGGATGTTAAGCCAAAT	N.A